Social network of firms, innovation and industrial performance

Johannes van Der Pol

université de BORDEAUX

THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR DE

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE ENTREPRISE, ÉCONOMIE ET SOCIÉTÉ N°42

SPÉCIALITÉ : SCIENCES ÉCONOMIQUES

Par Johannes VAN DER POL

Social network of firms, innovation and industrial performance

Sous la direction de :

Murat YILDIZOGLU et Andreas PYKA

Soutenue le 17 Novembre 2016

Membres du jury :

M. **LISSONI**, **Francesco**
Professeur des Universités, Université de Bordeaux, **Président du jury**
M. **PYKA**, **Andreas**
Professeur des Universités, University of Hohenheim, **Co-directeur de Thèse**
M. **VALLEE**, **Thomas**
Professeur des Universités, Institut d'Economie et de Management de Nantes, **Rapporteur**
M. **VERSPAGEN**, **Bart**
Professeur des Universités, Maastricht University, **Rapporteur**
M. **YILDIZOGLU**, **Murat**
Professeur des Universités, Université de Bordeaux, **Co-directeur de Thèse**

Titre : Réseau social des firmes, innovation et performance industrielle

Résumé : L'objectif de cette thèse est de répondre à trois questions principales ; comment expliquer et interpréter un réseau de collaboration, est-ce que des firmes avec une position particulière dans un réseau bénéficient d'une performance accrue et enfin, existe-t-il des structures de réseaux qui favorisent l'innovation ?
Pour répondre à ces questions, la thèse est organisée en trois parties. La première partie présente, dans un premier chapitre, une revue analytique de la littérature suivie d'un chapitre qui présente la théorie derrière une des méthodes d'analyse réseau utilisée dans cette thèse : les Exponential Random Graph Models (ERGM).
La seconde partie présente trois analyses empiriques. Le premier chapitre empirique analyse l'impact du cycle de vie de la technologie sur la dynamique du réseau de collaboration autour des composites structuraux en aéronautique. Les deux chapitres suivants se concentrent sur secteur aéronautique et le secteur des biotechnologies respectivement. L'objectif de ces chapitres est d'analyser la dynamique structurelle et d'identifier s'il existe un lien entre position dans le réseau et la performance de la firme.
La dernière partie cherche à identifier des structures de réseaux qui favorisent l'innovation. Un modèle à base d'agents (ABM) est proposé pour répondre à cette question.

Mots clés : Réseaux d'innovation ; ABM ; ERGM ; Performance ; Small world

---

Title : Social network of firms, innovation and industrial performance

Abstract : This thesis aims to answer three main questions ; how can one explain and interpret the structure of an innovation network, are there positions in a network which allow for an increased performance for firms and finally, are there network structures which favour innovation ? In order to answer these questions, the thesis is organised in three parts.
The first part presents, in a first chapter, an analytical review of the literature followed by a chapter presenting the theory behind one of the network analysis methods: Exponential Random Graph Models (ERGM).
The second part of the thesis presents three empirical analyses. The first empirical chapter analyses the impact of the life-cycle of the technology on the structural dynamics of the collaboration network for Structural Composite Materials. The following two chapters focus on two sectors, the aerospace and biotech sector. The aim of these chapters is to analyse the structural dynamics of collaboration networks as well as identifying a link between network position and firm performance.
The third and final part of this thesis searches for network structures which might favour innovation. An Agent-Based Model is used to answer this final question.

Keywords : Innovation networks ; ABM ; ERGM ; Small world ; Performance

---

GRETHA UMR CNRS 5113 – Université de Bordeaux
Avenue Léon Duguit, 33608 Pessac Cedex

L'université de Bordeaux n'entend ni approuver, ni désapprouver les opinions particulières émises dans cette thèse. Ces opinions sont considérées comme propres à leur auteur.

Ik draag dit werk op aan mijn opa,

“The greatest pleasure in life is doing what people say you cannot do”

– Walter Bagehot

Acknowledgements

First I would like to acknowledge the people who made this work possible: I would like to thank professor Yildizoğlu for introducing me to agent based modeling and accepting to supervise this thesis. His meticulous comments during the writing of this thesis and his scrupulousness have had a significant influence on the quality of this thesis. Working together on the ABM model has been a real education.
I am grateful to Professor Pyka for accepting to participate in this thesis with a late start. His considerations are greatly appreciated. I would also like to thank the members of the jury for accepting to be a part of the conclusion of this work. The interest they have shown by accepting is highly valued. This thesis is financed by IdEx Bordeaux, I am very thankful to them for choosing to finance this project and thus allowing me to bring it to completion.

This being a work about the transfer of knowledge and collaboration I see it fit to thank my own sources of knowledge and collaborators on this and other projects. In this light, the members of the GREThA have been available and willing to share their expertise, for this I am grateful.
Anissa, thank you for your advice on many of the questions around this thesis. I will try to remember not to be stupid.
Mathieu, thank you for allowing me to attend and present at many VIA-INNO meetings, the discussions have been invaluable. Introducing me to Damien Talbot to discuss the aerospace sector has been of great help. Working with your team has been a pleasure and I thank all the engineers of the VIA-INNO platform for their valuable input. In particular, Jean-Paul, your expertise on query building and database manipulation has been more than useful. Working together on our wacky ideas is a pleasure. Jean-Paul, Bernard and David, your knowledge of the aerospace sector and SCM has been very useful, thank you for sharing with me. Marina, your challenging questions have made the projects we worked

on that much more interesting.

Francesco, thank you for many useful discussions and opportunities, they have been incredibly useful.

Marie, Christophe, Vincent, thank you for many discussions and advice.

Nicolas, thank you for helping me acquire the much needed extra years of financial data.

I would also like to thank Olivier Baron, Emmanuelle Gabillon, Pascale Roux, Sebastien Rouillon, Marc-Alexandre Sénégas, Martin Zumpe for trusting me to teach their classes. I thank Caroline, Laurence and Sandrine for making the administrative chores that much easier to handle.

Last but not least I thank my family for their support through the years.

Bordeaux, November 17th 2016

Johannes van der Pol

Contents

List of Figures

List of Tables

Introduction

Introduction

The technological landscape evolves continuously. New technologies are discovered, others are recombined into new products. In this context of constant technological change, firms need to adapt in order to survive on the market. To ensure that firms will not fall behind their competitors, firms are required to continue researching and developing their products. However, with increasing complexity, the number of different technologies required to produce a product increases. Firms might face the fact that they do not possess the required knowledge to continue developing their product. The firm then has to decide whether to develop this knowledge in-house or to collaborate with another firm that has experience in the required field. Developing the knowledge themselves would be both time consuming and bear a high financial cost. Seeking a collaborator that already has part of the required knowledge would then seem like a viable strategy (Pyka, 2002). Combining the knowledge within firms allows for the discovery of new technologies or the improvements on the existing ones. In this light, collaboration has proven to be beneficial for the firm (McEvily and Marcus, 2005), for innovation (Kogut and Zander, 1992; Tsai, 2001) as well as survival and growth (Watson, 2007). As a result, in the past few decades the number of collaborations between firms has been steadily increasing (Saviotti, 2007; Tomasello et al., 2013).

As a result, in the past few decades the number of collaborations between firms has been steadily increasing (Saviotti, 2007; Tomasello et al., 2013). When we aggregate these collaborations in a certain context, a network is created. As such we can define a network around a specific technology, a sector, a geographical region and so forth. The main objective of this thesis is to analyze these networks, study how they evolve and how firms benefit from them. We work under the assumption that during collaboration, firms exchange knowledge. This exchange can be either voluntary (technology swap, training, license, employee discussions), or involuntarily (the simple act of observing how others work can

result in the copying of routines and information).

Collaboration hence allows firm to create new knowledge together while at the same time diffusion their preexisting knowledge to their collaborators. Knowledge can then flow through the network and be used by other firms in their innovation process.

This exchange of knowledge is however not always perfect (Arrow, 1962). Knowledge is one of the most important assets of the firm (Veugelers, 1998), as such it needs to protect this asset while trying to access knowledge held by other firms. Deciding how much knowledge a firm is willing to disclose is part of the strategy of each firm. In addition, the quality of the transfer itself depends upon the ability of the firms to send or absorb the knowledge they are exposed to. Firms exposed to knowledge that is too advanced might be unable to successfully integrate it into their R&D or production process. In this case the knowledge flow has no real impact. Firms exposed to knowledge that has already been integrated will find that their processes are not significantly influenced by the received knowledge. On top of this, if the firm who is supposed to send the knowledge is not efficient at transferring, the knowledge flows will have little to no impact either. Collaborations are hence a delicate matter. Firms can act as inhibitors or catalysts when is comes to knowledge diffusion. These aspects are not only important when we focus on the level of the firm, but also when we look at the larger picture. Given that firms can have more than one collaborator. Some of the knowledge they have gathered, combined with their own knowledge, can in turn be exchanged with other collaborators. From firm to firm, knowledge flows through the network, either slown down or accelerated by the different firms it passes through. The pattern connecting all the collaborators together, the structure of the network, then appears to have a major impact on the knowledge to which each firm is exposed.

In densely connected networks, knowledge is send many different paths which can speed up the diffusion to the rest of the network. In addition, in dense networks the average distance between firms is small and knowledge does not have to travel far to reach all other firms. On the other hand, in sparsely connected networks, the travel time is vastly increased. The presence of gatekeepers in the network can be harmful from the diffusion point of view. If the firm in this cluster-connecting position is inefficient in sending the acquired knowledge, this deprives part of the network from the knowledge flows. The structure of the network has hence an important role to play in the study of the flow of

knowledge between firms.

A first concern in this thesis will be the understanding of the structure of the network. In other words, we want to be able to describe the global structure and its dynamics, explain the different clusters that compose the network and shed light on the strategies of the firms in terms of partner choice. Understanding the structure is of interest not only because of the question of the flow of knowledge, but also because it provides insight into the organisation of the R&D processes of the firms. The network is an aggregation of collaborations and hence reflects the strategic decisions of the firms.

The idea that firms can benefit from knowledge held by other firms raises another question. If knowledge flows are important for the innovation process of the firm, then there should be a link between networking and performance of the firm. The second concern in this thesis relates to the performance of the firms. More precisely, are there positions inside the network that are more favorable in terms of knowledge flows ? In other words, is there a position in which firms are able to capture more knowledge and/or a position that allows firms to better integrate new knowledge ? If this is the case then these firms should benefit from a better performance than firms with less favorable positions. Do delve deeper into this, are there global network structures that favor the performance of firms ?

To sum up, we have three main questions: First, how can we explain and interpret the structure of a collaboration network ? Second, since different positions inside the network expose firms to different levels of knowledge flows and a more or less favorable environment, do firms with specific positions inside the network outperform those with a less favorable position ? And finally, are there network structures that are more favorable for innovation ?

To answer these questions, a review of the literature around the link between the structure of the network and the performance of the network is required. To take into accounts the answers already given to these questions, an analytical literature review may help. Consequently, in a first chapter on the manner in which firms exchange knowledge. The chapter will review how different types of knowledge can diffuse through different types of interactions between firms. A link is then made between this diffusion and the manner in which it is impacted by the structure of the network. In the second chapter,

before switching to the empirical analyses, a methodology for the identification of firm level link creation strategies is presented. This statistical method is, in my opinion, highly useful for understanding the structure of networks. This method, Exponential Random Graph Models, is based on logistic regressions. The main modification added is the lifting of the hypothesis of independence of observations. In classic regression we make the hypothesis that all observations are independent. In the case of networks, the dependence of observations is an important factor. The emergence of a link can very well depend upon the structure of the network before the link was created. For instance, if a firms has two collaborators, the probability that these two will come in contact is higher than two other firms chosen at random. The probabilities are hence dependent upon the existing network. ERG models are able to account for these dependencies and are hence the perfect method for explaining the structure of a network from the micro level to the macro level.

In the second part of the thesis, we will use this empirical method for understanding the dynamics of networks and the performance of firms in specific sectors, particularly relevant from the innovation networks perspective. The first analysis will focus mainly on the first question. It aims at identifying a determining factor in the structuring of innovation networks. Since the structure taken by a network is highly impacted by its context. Factors such as industry (Salavisa et al., 2012), types of actors included (Nieto and Santamaría, 2007) as well as geography (McKelvey et al., 2003) have shown to have an important impact. However, the role played by the technology life-cycle is still mostly unexplored (Stolwijk et al., 2013). We could expect that different stages of the technology life-cycle require collaborations with different firms or research institutes. A research stage calls for basic research and knowledge about fundamental technologies, while the development phase requires collaborations with firms that have a more applied approach to the technology.

In order to check this hypothesis I chose a collaboration network around one specific technology: structural composite materials in aeronautics. The impact of the life-cycle of the technology would not be visible at another level of analysis .

The aim of this third chapter is to extend the existing literature on network formation by showing how the life-cycle of the technology impacts the formation of the collaboration network. In addition, we extend the existing literature on technology life-cycles by showing that networks can be used to identify the different stages of the life-cycle.

The following two chapters continue to focus on the question of network formation but also aim at answering the first question on performance. The aim of these chapters is to analyze the network structure to understand how knowledge is created and what mechanisms drive link formation between firms in the different sectors. Answering this question starts with understanding how firms pick their collaborators and how this shapes the network. The structure of networks can be analyzed at different levels, each level providing information about the structure. At the highest level, a network is analyzed as an entity on its own. The global structure of the network is analyzed. This first level of analysis provides the identification of the most prominent actors in the network, the presence of clusters, gatekeepers. It allows us to understand if the network is an interconnection of well defined clusters, one large cluster or a very sparsely connected network. To help with the analysis of networks, different methods exist that allow for testing if a network has a particular structure of which the characteristics are well defined and well understood. Small worlds for instance, are defined by a high level of clustering and a low average distance between the firms in the network. These networks are positioned (from a structural point of view) between completely random and regular networks. Random networks have a screwed degree distribution while in a regular network, all firms have the same number of links. Small world have been observed empirically in social networks as well as economic networks, and have been included in many theoretical models because of their empirical relevance. These studies have shown that small worlds are efficient for the diffusion of knowledge thanks to the low average distance and high clustering. The short average distance ensures that the knowledge diffuses quickly through the network. The high levels of clustering are often signs of dense local interactions between agents. Think of groups of friends, firms collaborating in a production chain, geographically close agents of any kind. These local interactions foster the generation of knowledge which then rapidly spreads through other local clusters. Another example of a widely known network structure is the core-periphery structure. The latter is a structure which is related to Pareto's law. A small portion of the nodes in the network have many links while a large portion has only a few links. Core-periphery structure have been found in many different settings. The idea is that a couple of large firms with a large number of collaborators are densely interconnected. The smaller firms are connected to a small fraction of the larger firms, and hence placed at the periphery of the network. This gives a particular structure to the network, a couple of

interconnected stars surrounded by many, sparsely connected nodes. These networks are close to the rich get richer principle, take the example of citations on publications. Papers with a higher number of citations are more read than papers with less citations, as a results the probability that they will gather additional citations is higher. The same idea applies to collaborations, large firms can sustain and might attract more collaborations. Their degree in the network is more likely to increase than that of a small, emerging firm.

These network structures have their relevance in different settings. For instance, a production chain as the aircraft industry would be a network build from a few highly connected firms (Airbus, Dassault, Thales) and many smaller suppliers with less collaborations. It would hence make sense to check for the presence of this type of structure. In the case of more atomized sectors such as the Biotech sector, we would expect there to be a more homogenous distribution of the number of links between firms. Considering the high level of competition and the race for innovation, we would expect firms to give high importance to social links and trust in the choice of their collaborators. We would hence expect to observe a small world structure. The global network structure is the result of an interconnection of smaller networks, or clusters. Identifying the different clusters that shape the global network structure is a step towards understanding why the network has the shape that it has. This level of analysis between the macro vision of the global network and the micro vision at the level of the firm is an important step into understanding the structure of a network. As I will show in the forth and fifth chapters, the identification of clusters results in well defined clusters that can easily be explained by economic factors. In the case of the aerospace industry, each cluster represents a different part of an airplane. The production chain organization of the sector is clearly visible. In the case of the Biotech sector however, we observe a clear distinction of market tiers in the network. Once the structure of the network is clear, the question of performance is raised. According to their position in the network, firms do not benefit from the same knowledge flows. Some positions, gatekeepers, central firms, are more favorable than others. The question then is: do firms with a specific position in the network perform better than other firms ? Of course this question cannot be answered by taking into account only structural elements such as centrality and density. The neighborhood of the firm is a factor that has to be taken into account. The diversity of knowledge in the neighborhood of the firm as well as the experience in terms of collaborations are expected to impact performance. In addition, the

absorption capacity of the firm defines how the firm is able to absorb the knowledge it is exposed to. The variables, in addition to the structural ones, are used in a panel regression in order to explore the link between position in the network and firms' performance. The indicator used for the performance is the Return On Assets (ROA), the latter was retained since it has a broad definition of the assets of a firm and hence contains the financial value attributed to any knowledge the firm may have (in the form of patents, licences etc.).
Answering the final question, can we identify network structures that are more favorable for innovation, is close to impossible with empirical data. There is no relevant data that would allow the measuring of performance at the network level. Indeed, even if we can compute indicators, there would be no way to compare that to a control of firms not evolving in a network. Or at least there would be no representative sample. At least in the data used in this thesis, all large corporations were present in the network, and hence none can be used a control sample. Because of this lack of data and the difficulty of computing relevant indicators such as consumer surplus, a theoretical approach is appropriate for answering this question. Since the value of networks resides in the heterogeneity of agents, an Agent-Based approach was chosen. This method allows the modeling of a network in which each firm will have its own productivity, production, technological neighborhood, market-shares and so-on. This provides a more realistic model than those based on a representative firm. The final chapter of this thesis hence presents an agent-based model of innovation networks. The aim of the model is to identify the impact of the structure of the network on the performance of innovating firms. Based on elements identified in the analytical review and the empirical chapters, the model has a particular focus on the way firms include knowledge flows into their R&D process. The model is not an innovation diffusion model. Firms are able to provide their own R&D and investment strategies in order to adapt to the evolutions of the market.

We will proceed to the first chapter of the thesis, which will present what we know about innovation networks and performance. This will be presented in the form of an analytical review of the literature.

Part I

Part A: On the shoulders of giants

Chapter 1

Social interactions between innovating firms

"You want weapons? We're in a library. Books are the best weapons in the world. This room's the greatest arsenal we could have. Arm yourself!"– The tenth Doctor

1.1 Introduction

Innovations are the driving force behind growing economies and prosperous firms. Achieving innovation is hence at the center of any business strategy. Researchers since Schumpeter have focused on the role played by technological progress in economic growth. It has since been identified as one of the decisive factors enabling continuous economic growth.

When looking at this question from a micro-economic point of view we can ask how this technological progress is achieved, i.e how do firms innovate. Broadly speaking, innovations are achieved through the recombination of existing technologies and ideas. This process of recombination of knowledge results in a complexification and cross-fertilisation of different technological domains. The increase in complexity of the technologies used in the innovation process have as a consequence that firms are no longer able to master all the required knowledge. Accessing external sources of knowledge then becomes part of the innovation process. Firms are incited to go beyond their bounds to access knowledge held within other firms (Hagedoorn, 1996; Narula and Hagedoorn, 1999). The knowledge held within the firm becomes its main asset (Penrose, 1959). In this Knowledge Based View (KBV) of the firm (Penrose, 1959), firms have to protect their own knowledge while

gaining access to new knowledge held by other firms.

The need for firms to access new knowledge has become increasingly important (Duysters et al., 1999), making collaborations more widespread, changing the business landscape by a profound reconsideration of strategic decisions. Hagedoorn (2002) shows that collaborations have been steadily increasing in the 1980's, a trend that has continued ever since (Nesta, 2005).

The aggregation of these collaborations, whether at the level of a single technology, a sector, region etc., results in a collaboration network. The latter is viewed as an interconnection of collaborating firms with a common goal. A network is however much more than a mere sum of its parts. Each actor in a network influences, in one way or another, any other actor in the same network and the network impacts in return the performance of the firm.

The aim of this chapter is to study how innovating firms can be influenced by their network and how, in turn, they can shape the network. The document is organized as follows; The first section will explain the functioning of networks and how they emerge. Section two will analyze the impact of the network on the R&D process of the firm. A third section reviews how knowledge diffuses through the network while the final section studies how these flows impact the efficiency of the network.

1.2 The genesis and basic mechanics of innovation networks

When compared to working alone or in bilateral cooperations, networks present numerous advantages for firms. Networks are designed for long term interactions that go beyond the scope of a single project, firms interact continuously through time on different projects (at the same time or one after the other) allowing for social links to become stronger and for firms to learn from each other.

Bilateral cooperations are known to have a low success rate (Barringer and Harrison, 2000; Masrurul et al., 2012), reasons are suspected to be (mostly) diverging interests and managerial disputes. Networks are less prone to these risks since they are build with a common goal, creating all the more incentives for firms to invest completely in the project. Instead of collaborating in order to gain access to knowledge, firms have the option of

buying information or knowledge on a market. The market for knowledge raises a certain number of problems. The market for information is indeed imperfect, the uncertainty about the quality or even the source of the information making it difficult to put a price on information. If the buyer on a market "knew enough about the information, he would know the information himself" (Arrow (1962), p.946), the intrinsic risk that goes hand in hand with information creates imperfections in the market. Firms then cooperate and share their knowledge in order to overcome these imperfections. The social links that emerge from these cooperations allow for reputations to flow in the network resulting in an auto-regulatory environment that replaces market hierarchies. Networks are hence a stable form of industrial organization. Podolny (2001) summarizes the benefits of networks for participating firms by the use of two metaphors: tubes and prisms. Tuber refer to the potential of flow between connected firms, while prisms are a reference to the reputation that flows through the network.

1.2.1 Networks as an interconnection of tubes

Tubes refer to the links between firms. When firms cooperate a link is created between them, allowing for an exchange of knowledge, funds or any other resource they might be willing to exchange. Accessing other firm's resources is of vital importance for firms since with growing complexities in new technologies a firm alone can no longer master all the technologies needed for the production of a single product. Accessing different knowledge sources is thus beneficial for the firm (McEvily and Marcus, 2005), for innovation (Kogut and Zander, 1992; Tsai, 2001) as well as survival and growth

McEvily and Marcus (2005) show, by studying joint problem solving projects between firms and suppliers, that learning from a diversity of partners increases the competitive advantage of the firm on the market. They argue that firms acquire capabilities through exchange more than alone, the better the relationship between the firm and the supplier the easier the transfer of capabilities. This transfer is important because technologies will reach a point where their returns become decreasing (Kogut and Zander, 1992). The authors argue that because of this decrease in returns, the firms must access new capabilities. Learning new capabilities or even finding them is difficult. Firms embed themselves into a way of organizing their production, they will have to go out of their comfort zone in order

to achieve better returns, learning from other firms is one possible solution. Tsai (2001) supports the view that in a network firms can learn from each other and hence increase their innovative performance but does point out that the success of such transfer depends upon the ability of the firm to absorb the knowledge (as proposed initially by Cohen and Levinthal (1990)). The influence of networking on industrial performance should thus be visible and Watson showed that it is (Watson, 2007). Through a survey of Australian firms he was able to find a significant influence of networking activity on the survival of the firm, growth however was less influenced by network activity. Growth is shown to be influenced by collaborations by Szulanski (1996) and Duanmu and Fai (2007). The latter showed that Chinese suppliers learn from the routines of multinational firms significantly increasing their R&D productivity. Networks hence allow firms to interact and influence each other over time by the exchange of knowledge.

Knowledge is a powerful motivation for firm cooperation. There are however other motivations. Table 1.1 gives a list of different motivations that go beyond the benefits of knowledge exchange.

Other resources might be interesting to a firm, especially when these resources are rare. We can take the example of the large hadron collider. A network might emerge for the purpose of the exploitation of such a resource, to regulate its use and creation.

Firms can decide to cooperate with other firms for internalization purposes. Some firm might be brought to consider cooperations because it knows that other firms might profit from it's innovative activity without it's consent. In order to avoid being victim of this externality the firm can cooperate and hence internalize the externality.

Firms can also decide to enter a network for the aim of the network, for a specific cause. Firms in an industry with high pollution can decide to enter a network for the development of cleaner technologies purely for ecological reasons. It is however possible that this interest for cleaner technologies is motivated by new government regulations. This is the case in the aerospace sector for example, where cleaner engines are required by the european commission. Firms then cooperate in order to make this technology as efficient as possible. The advantage of this kind of cooperation resides in the fact that the technology cannot be used for a competitive advantage by any of the firms (or consortium). This facilitates the emergence of industry standards since the technology is developed by a

large number of firms which can come from all stages in a supply chain. This means that during cooperations firms will be able to be more aware of the different requirements and problems they might face. In addition, they can find common solutions increasing the efficiency of the technology. This also implies that during cooperation firms learn about each other's technology resulting in an increase in efficiency of the innovation process. Not only do firms learn about the technologies of other firms, they also learn about the trustworthiness and value of a collaborator. The network is not only a catalyst for the exchange of knowledge relative to technologies but also for reputational aspects. From this point of view networks act as prisms.

Aim	Definition
Specific goal of the network	Networks can be created for a specific goal; greener technologies, healthier products, computer standards etc. Firms might decide to enter a network purely because it believes in the cause of the network.
Internalize externalities	A firm can anticipate that other firms will benefit from its efforts. The creation of a network will allow the firm to limit the free rider effect.
Standards and labels	Using standards is beneficial for a firm since it allows to cut cost and might result in a larger user base (for electronics: wi-fi, usb). ISO standards , controlled designations of origin and other labels.
Access to rare resources	Firms might require the use of a specific resource (technology,natural resource etc.). A network of entities might manage the resource, in which case a firm can decide to join the network with the sole purpose of gaining access to the resource in question.
Access to a new market	Entering a new market is risky, collaborations with incumbent firms can reduce the risk factor since the firms can share their knowledge of the market.
Access to complementary knowledge.	Firms might find that it is more cost-efficient for firms to collaborate in order to use knowledge mastered by other firms rather than invest in R&D in order to master the same knowledge.
Access to funding	Policy makers can put conditions on the distribution of funding for R&D which can include collaborations.

Table 1.1: *Motivations for collaborating*

Collaborations between firms allow for more than just the exchange of ressources between firms. Since collaborations are risky endeavors, and prone to failure (Masrurul et al., 2012), the search process for new collaborators is influenced by the reputation and

trust of prospects.

1.2.2 Prisms and the diffusion of reputation

Repeated cooperations between firms allows trust to form. Information about the trustworthiness of agents then flows through the network (mainly by social contacts). A collaborator that slows down projects or behaves as a free-rider should at all cost be avoided. The choice of partner is both difficult and of paramount importance. Problems of moral hazard and adverse selection can be reduced by basing partner choice on repetitional effects. Un-cooperative behavior will result in firms having a bad reputation, keeping them from collaborating again in the future. Increased trust allows for a decrease in the failure of cooperations as shown by Zaheer et al. (1998). The authors show that there is even a direct link between trust and the performance of the firms (as measured by competitive price, quality of goods delivered and respect of deadlines). These results were obtained by studying dyadic exchange relationships of electrical equipment manufacturers. Even though the object studied was not a network this is only because the authors did not define their cooperations as a network. There might not be a common goal for the actors here but their reputation can still flow through the social network of the firms with whom they work. As such the social network of the firms influences the performance of the firms through the creation of trust that result from dyadic cooperations.

Each agent inside the network sheds light on the other agents, either illuminating their positive reputation and capabilities or their un-cooperative behavior. From a dynamic point of view, the network allows firms to better select their partners by increasing available information about their knowledge. The flow of this information increases allows firms to better understand the functioning of other firms (either their routines or the employees they have to cooperate with). The more firms cooperate, the better they are able to exchange and combine their knowledge (Cowan and Jonard, 2007). Gulati (1995) shows that firms look for partners with whom they have a high cognitive embeddedness and who can provide them with new knowledge. The prism metaphor explains that in a network it is easier for firms to identify which potential partners possess the knowledge they are looking for and hence increase their efficiency.

This shows us that there are different dimensions to a network: we can find social links, informal links, contractual links and even invisible links. These links all play an important

role in the innovation process as we will see later in this document.

Leung (2013) notes that networks resemble sponges, they absorb information from its components, recombines these elements to create new knowledge which is send to the actors in the network, the process can then start over again. The network hence evolves, and not only because of the knowledge it created but also by the structure. The agents in the network rewire themselves to cooperate with firms with a better reputation or better knowledge enhancing once again the performance of the network. The strength of the network resides in its ability to evolve over time, getting rid of bad elements and innovate continuously by sending relevant knowledge to the firms composing it.

Networks can however have a negative impact on the firms. Pippel (2013) points out three potential technology-push disadvantages of networks, they mostly stem from what we initially considered to be advantages.

When a firms has identified indispensable resources for its R&D project that it does not possess itself, a partner needs to be found. The search for a fitting partner is costly, not only is it time consuming, gathering information about trust and quality of a potential partner is difficult to obtain and hard to verify. There is hence an inherent risk in choosing a partner.The network helps to reduce the cost and the risk. Working in a network implies that firms know each other and communicate on a regular basis, this allows for them to discuss behavior of other firms and judge the quality of their work.
This means that a firm searching for a new partner will be inclined to activate it's social network for this search rather than take the risk of cooperating with a firm it knows little about. As a result networks tend to become more locally clustered, phenomenon that is observed empirically (Hanaki et al., 2010). This phenomenon is amplified by social pressure. Firms that tend to cooperate often have close social ties, asking another firm rather than a socially close one is decision that becomes increasingly difficult, resulting in a specific type of lock-in, a social lock-in. Instead of choosing a new partner, (with potentially new, more valuable knowledge a firm will continue to cooperate with the same firms. In fine the social lock-in can result in a technology lock-in because of the absence of new knowledge in the innovation process.

Pippel (2013) also points out that knowledge flows are difficult to control and some flows can be involuntary. Even in networks, specific knowledge still is a competitive advantage. Through cooperations knowledge can flow further than an initial dyadic cooperation resulting in a possible decrease of the competitive advantage.
Firms thus have to make sure they protect their knowledge and what they send throughout the network. This will greatly depend on the type of cooperations a firm has as the next section will show.

Networks are build up from tubes between agents.These tubes allow firms to access resources inside other firms that are needed for the accomplishment of their R&D process. Alongside the tubes that connect the agents, networks act as prisms that allow information on firms to diffuse inside the network. Firms will be able to chose partners based on their reputation. It is not always possible for firms to know which firm detains which resource, information about these resources flow in a network and will allow for firms to know with whom they can cooperate.

We have shown here the general conceptualization of networks and their functioning with a specific aim on innovation networks. The time has come to go into more detail on the functioning of the network by looking into the manner in which the different flows inside the network influence the R&D process of the firm and enhance their performance.

1.3 Networks and their influence on the innovation process of the firm

Dosi (2000) gives a detailed description of the innovation process that underlies the majority of innovations[1]. In his paper he reminds us that there are two theories describing the innovation process. These two theories, demand-pull and technology-push, explain where ideas come from and are hence the starting point of the innovation process.
What we will recall from these theories (for our purpose) is that one theory suggests that the market sends information about it's needs to the firms and that this information

[1]With the exception of innovations that are created purely by accident

should be used by the firms for new ideas. The other (technology-push) suggests that firms innovate without consulting the market, pushing an innovation without there being demand. We can illustrate the difference with the example of the iPad. The iPad was created by a company without there even being a market for it. The firm created the product and pushed it on the market. Hybrid cars on the other hand are the result of a reply to a consumer demand for the reduction of cleaner cars and a reduction of fuel consumption.

1.3.1 Demand-pull theory

Demand-pull theory suggests that the consumers reveal their preferences regarding innovation by their behavior on the market. If goods with certain characteristics have a higher demand, firms should include these characteristics in their new products.
The question we have to answer here is how does the network influence this first step of the process. The demand-pull theory suggests that firms have information about consumer preferences. It should be clear that this is valuable information since innovations based on this information have a high probability of finding demand on the market. Firms that posses this valuable information do still have an incentive to collaborate. A single firm is unable to master all the technologies needed to create a product. Working with firms that have the reputation for providing high quality products can only increase the quality of the end product that they develop together. Disclosing their information might not seem as a viable decision in the short run, because profits will be shared and the competing[2] firms in the network might become more capable. It is however a viable strategy in the long run. Sharing valuable information allows a firms to show it's value as a collaborator.
The same reasoning applies for firms sharing their productive capacities. By showing what one can do to other firms one ensures future cooperations and thus opportunities to learn and access more information. This information is valuable in its own right because it can result in an increase inefficiency for the firms.

[2]Firms that compete on the same market

1.3.2 Technology-push

Technology-push implies that firms create innovations without knowing if there will be a demand for the product on the market.

Cooperations between numerous firms allows an exchange of ideas an expertise in various domains. If no information on demand is available, the risk that an innovation will find demand is higher. The reason why firms still engage in technology push innovations lies in an information asymmetry between the market and the firms. New discoveries (by universities or labs for example) can open the way for a large number of new products that consumers had not yet imagined to be possible.

Once again it is the diffusion of information that allows for ideas to emerge. Only experts can see the potential of new discoveries and estimate their impact in their industry. Allowing this information to spread as widely as possible is hence positive not only for the consumers but also for the agents developing the technology. A network provides agents with contacts whom they can trust (their reputation has been established by their previous accomplishments in or outside the network), which allows firms to bypass the market with all it's imperfections. The discussions between the agents, and hence the combination of their expertise will allow for a development under optimal conditions. By contrast, a firm working on it's own would have to acquire abilities by itself, requiring important investments. A network allows direct access to agents who can share their experience and reduce the risk of failure.

One special case of a technology-push is worth mentioning here, the development of a new industry standard.

Standards in an industry influence the efficiency of both the production process and the innovation process. In any cooperation time is needed for firms to adjust to their work methods and organize the compatibility of technologies. The use of an industry standard allows these efforts to be drastically reduced and the efficiency of the cooperation increased. In the french aerospace sector for example the whole supply chain uses software designed for the sector. The software makes sure that all parts are in stock. As soon a an order is placed it immediately updates stocks and orders for all the other firms in the sector.

As we discussed before, the instauration of such a standard can be a motivation for network creation, showing the importance of standards in an industry. The advantage of the

development of a technology in a network resides in the fact that it will be adopted by all firms. If standards emerge from a market there is a possibility that we observe competition between standards resulting in a loss of efficiency until one of the two competing standards wins.

Until now we have mostly discussed knowledge exchange between firms. Firms are however not the only source of knowledge. Research institutions such as laboratories and universities can be the epicenter of new technologies that might initiate new technologies and products. Indeed, it is in these research centered institutions (RIs) that fundamental research is developed that will define the technologies and materials of tomorrow. The research provided by these institutions results mostly in codified knowledge (by patents or publications). The difference with the knowledge provided by firms resides in the application of the technology. Firms develop their technologies and routines with the specific aim of making them operational inside the firm. The research institutions (RIs) do not have this aim, they are only motivated by the performance of the technology they are developing. As such, if no cooperations exists firms would not be able to efficiently exploit new technologies (or not at all).
Networks then allow firms to help RIs to find an application for their developed technologies. The objects of the exchange hence are far beyond the transfer of tacit information, the firms have to develop new routines and methods with the knowledge provided by the RIs.

The network then influences greatly the generation of new ideas at the beginning of the innovation process and how they are developed. Firms are no longer shielded from technological progress in other firms nor from the development of fundamental research (through cooperations with research institutes). This makes it easier for other firms to learn who masters which technology in the network and hence facilitates the second step of the innovation process.

1.3.3 The search for resources

The second step in the innovation process is the identification of resources needed for the development of the technology. Resources such as new technologies, patents,

machines, expertise and know-how are identified. Taking into account the cumulative nature of knowledge, firms that wish to innovate need to master the existing technologies that incorporate this knowledge (Cowan and Jonard, 2007).

The identification of firms that master the needed technologies is not easy. Indeed, it is not simple to find out which firm uses a certain technology, who hold vital patent that need to be licensed etc. For firms in network this step is greatly simplified.

The search for partners will be accomplished in part through the "prisms" of the network, as time flies by and collaborations begin and end more is known about which firm knows how to do what. This information will flow through the network and should converge to a perfect information scenario. This is only possible of course if no new firms enter the network which is not necessarily the case.

A firm working on its own would have much more difficulty and also be confronted with hostility from the other firms who might not want to diffuse any of this information.

For example, Airbus wants to use composite materials instead of aluminum in order to reduce the overall weight of its aircrafts. The technology is relatively new and it has no in-house knowledge on the subject. By searching through its collaborative network it has been able to identify swiftly which of its suppliers was already working with the technology and which firms owned strategic patents on the technology that needed to to be licensed for the development.

1.3.4 The development stage

The research for resources being over, the partners with whom a firm is going to cooperate have been selected allowing for the research and development stage to start. Tubes are now created between the firms which opens the possibility for exchange between the agents. The firms start to work together and will hence be able to observe the inside of the firm. Hence, more than information, knowledge can be exchanged during this stage. More specifically, two basic types of knowledge can be transferred through the tubes: Routines and technology specific knowledge.

Routines The concept of routines was first introduced by Nelson and Winter (1982) and refers to the way a firm is organized and accomplishes its tasks. Firms improve their routines over time, and can share their experience with other firms. For example Duanmu and Fai (2007) showed that Chinese firms, by observing how multinational firms organized their R&D, were able to better organize their own R&D procedures resulting in higher productivity. The links in their study are vertical but horizontally the exchange is possible as well. Even competitors can learn from each other's routines.

Networks allow for firms to exchange best practices (Szulanski, 1996) or even encourage it since it will allow for a smoother cooperation between firms. Converging work methods will allow for agents to speak a common language in the development process and avoid lost time due to inefficient organizational aspects. The firm that has the best practice will benefit just as much as the other firms from the exchange of routines. Firms are hence able to learn from each other and implement best practices that increase the efficiency of the organization of the production (and/or R&D) process. The network influences the R&D process by the means of gaining access to best practices from other firms which would not be willing to exchange this knowledge in a more competitive environment.

Technology specific knowledge Technology specific knowledge is the know-how that is needed to operate a machine or the knowledge to understand how a technology works. In the case of a network (or a simple cooperation) firms might benefit from both mastering a technology. Some technologies are however not easily operational without the help of an expert (Comin and Hobijn, 2004). Think for example of a firm that masters composite materials and uses it to make wings for an airplane. If it has to work with a firm that wants to use the materials to make nacelles, then for the sake of having a motor that is correctly fastened to the wing, the first one might want to teach the second one about composite materials. This exchange of knowledge is beneficial for both parties because it increases the quality of the components. Considering that firms in networks work towards a common goal they are concerned with the quality of the objective the network and not only they part they supply.

Networks then, incite firms to exchange knowledge and broaden the range of the technologies they master in the process. In the case of a technology that was not unknown

by the firm the exchange can have implemented a better knowledge of the technology. For example firms could now be aware of other applications, or have more ease in the resolution of problems with regards to the technology. Working in a network setting allows firms to learn new technologies, increase they level of comprehension or even solve existing problems.

The network influences the R&D process of the firm in each of the stages of the process. Networks make it easier to access knowledge and information that allows the firms to be more efficient in the choices it makes but also in the development of the technology that has been chosen.

However, the diffusion of knowledge as we just described is not an easy process and is subject to many factors that influence the efficiency of the learning process through a network. The efficiency of the transfer will depend on the channel through which it has to travel but also on the type of knowledge.

1.4 The diffusion of knowledge in a network

1.4.1 Forms knowledge may take and the influence on its diffusion

So far we have identified two types of knowledge that were of vital importance in the R&D process: Routines and Technology Specific Knowledge (TSK). These terms are general, they contain in fact various types of knowledge with different characteristics. These characteristics define the transferability of knowledge and hence how it flows through the network. Polanyi (1966) and Nelson (1990) made a first distinction between different types of knowledge. Polanyi defends the opposition between implicit and explicit knowledge, where explicit knowledge is a physical aspect of a technology (machine) and the know-how to operate the machine is referred to as implicit knowledge. It is obvious that the machine is in many cases easy to transfer from one firm to another, yet acquiring the know-how to operate a machine takes more time and effort.

Nelson draws our attention to the fact that implicit knowledge itself can be decomposed into two major components, a 'private' and a 'public' component. The private component

refers to technology specific knowledge, i.e knowledge that is only of use for a certain technology (system adjustments or specific problem solving).
The other component is the public component, which englobes all general knowledge used to render the machine operational. An example of general knowledge could be that the machine should be plugged in, and in order to make it work one has to turn it on. This distinction is still valid if we take into consideration information without a physical component. For example, if one wishes to compute the maxima of a non linear system, one will need basic algebra skills involved that would be considered to be common knowledge, but there are also other more complex methods necessary that are more specific to the task at hand, which would be the private component (the programming of algorithms). The use of a machine is only one side of the story. Nonaka (1991) makes a similar observation but uses the terms "explicit" knowledge and "tacit" knowledge. This distinction has lead to what is nowadays referred to as the "Knowledge Based Vision" of the firm (Penrose, 1959). Nonaka uses the analogy of an engineer who tries to develop a new bread making machine. The engineer tries to learn how to knead the bread with one of the top bakers of the country, and discovers that the knowledge the baker has about kneading is tacit. It has no physical form and thus the only way to learn it is through practice. While the two work together the tacit knowledge of the baker is transformed into explicit knowledge for the engineer (Nonaka, 1991), who can then transform his knowledge into a machine, that can be easily be transferred from company to company. This becomes even more obvious in Rogers' theory of diffusion (Rogers, 1982) who makes the distinction between hardware and software. Hardware is useless without software, and more importantly hardware can only be optimally used if the software is efficient.
The underlying hypothesis here is that the knowledge of a firm is held by the employees of the firm. Some authors have however noted that in the knowledge base vision of the firm, the firm as an entity can also posses know-how which can be tacit Kogut and Zander (1992). This vision relates to Nelson and Winter's theory of routines, in which routines are not only established by the employees but also by the firm.

In fine, two distinction should be retained for the analysis of diffusion of information, whether knowledge is tacit and whether it can be codified. Codified knowledge is the possibility for information to be passed down on a written support (patents, manuals, pub-

lications etc.). This allows us to organize all information into two categories. Information that has to be exchanged by face-to-face interaction and information that can be transferred directly between two agents without face-to-face interactions being a requirement.
Clearly, the flow of knowledge is faster in the case of codified information. The quality of the transfer will however depend on the absorption capacity of the firm.
What these descriptions show is that the tacit, implicit, private or codified dimension of knowledge need specific conditions under which they can be transferred. According to the type of cooperation chosen by firms certain types of information may transfer, other may not. Let us hence take a look at the different channels that may be created between firms and the objects they might carry.

1.4.2 The diversity of channels

Channels between firms can take a large variety of forms, from a simple discussion between employees to a joint venture or even a buyout. Table 1.2 gives a number of "tubes" that allow for knowledge exchange between firms. The differentiation is important because each tube (or channel) allows for different types of knowledge to flows through the channel and defines what firms may learn from one another. For instance, through a social interaction between employees only general problem solving can be transferred (Breschi and Lissoni, 2001), the influence on the productivity of the firm is only marginal.

Some links allow for bi-directional knowledge flow to occur, i.e Alliances, joint ventures, technology swaps and vertical links. These links will allow for the most valuable knowledge to flow. Indeed, these interaction are long term and hence allow for repeated interactions allowing tacit information to flow.
In the case of a buyer-supplier link the buyer might send employees to the supplier to teach them how to build parts up to their standards (cf. Airbus). This will have a significant impact on the performance of the firm, it will create a signal of quality and increase demand. Nike's supplier Mizuno was for example able to launch its own brand worldwide after learning from Nike.
Other links only allow for a unidirectional flow, Spin-outs, buy-outs, IP-transfers, R&D contracts and employee mobility. Even though these links allow tacit information flow,

Tube	Definition
License	A firm pays another for the use of a patented technology
Joint Venture	Two or more firms create a new firm for a specific purpose
Alliance	The pooling of resources by several firms
Social	Any contact between employees of a firm that may take place inside or outside the firm
Spin-Out	Employees of a firm create their own company
Externalities	Knowledge flows between firms or employees
Buyout	One firm takes ownership of another firm
Supplier	One firm supplies an intermediary good to another firm
OEM	One firm has an exclusive contract with another firm for the trade of an intermediary good
Technology swap	Two or more firms allow each-other to use a technology
IP transfer	The passing of hands of Intellectual property
R&D contract	One firm is contracted to perform R&D for another
Minority Investment	One firm buys less than 50% of the shares of another firm
Employee mobility	The knowledge stock of a firm is held by the employees of the firm. When employees switch firms they take part of the knowledge with them.

Table 1.2: *Diversity of tubes*

the flow only goes from one firm to another, there is no counter part. This is hence less valuable for the efficiency of the network as a whole. Innovating ideas and technologies will flow slower through the network.

Table 1.3, shows the different tubes and the types of knowledge they might carry.

	Alliance	Licence	JV ou RJV	Social	Spin-out	Externalities	Buy-out	Vertical	OEM	Technology swap	Ip-transfer	R&D contract	Employee mobility	Investment
Routines	■		■		■		■	■	■				■	■
TSK, explicit	■		■				■	■	■	■		■		
TSK, implicit, tacit	■		■		■		■	■	■	■	■		■	
TSK, implicit, codified	■	■	■		■		■	■	■	■	■	■	■	
Solutions	■		■	■	■	■	■	■	■			■	■	

Table 1.3: *Tubes and the objects they may carry*

The choice of the type of link depends on the knowledge flow that might result from the interaction. The structure of the network that results from these decision is hence partly defined by the type of knowledge pursued by the firms in the network.

This means that according to the needs of the sector in which the network evolves it is possible to find actors that generate fundamental knowledge. In technology intensive sectors such as biotech and aeronautics, research conducted by universities and laboratories is the only mechanism that allows for radical innovation to occur. The research provided by the Research Institutions (RI) cannot be completed inside firms, they lack not only the knowledge but often enough the research provided by RIs is void of any marketable application. Cooperations between firms and RIs allows for firms to help RIs to orient their research and market it efficiently.

I will make a second distinction between channels that occur between firms and channels that involve RIs. The RIs typically provide codified, scientific knowledge that has to find an application. This application is in most cases provided by firms. Where RIs are involved, the transfer of knowledge is relates to fundamental knowledge which has

a general nature. Transfers between firms are mostly transfers of technology specific knowledge. The transfer of knowledge between RIs and firms is different from knowledge flows between firms.

Channels between firms

Firms can interact in a variety of ways as shown in table 1.2. Not all of these tubes allow for the same type of knowledge flows. Some of the tubes allow for bilateral knowledge flows (RJV, technology swaps) while other only allow for knowledge to flow in one direction. The distinction is important for strategic purposes. Firms aim to protect their knowledge base, unilateral transfers protect firms from losing a competitive edge. Bilateral flows imply that firms need to share some of their knowledge base which can result in losing part of their advantage. In addition, when collaboration results in the requirement of receiving knowledge, the firms become dependent on the other firm. The choice of the type of collaboration is hence of vital importance.
IP transfers for instance are simply the transfer of a patent from one firm to another, the direction of the flow of knowledge unidirectional. A spin-out keeps the same spirit since a new firm is created while taking knowledge from another firm, without sending anything back. The case of R&D contracts is more complexe since the direction depends upon the nature of the contract. Research that is accomplished for the account of another firm cannot be considered to be bilateral knowledge flows. One of the firms creates the knowledge while the other receives it. The receiving firm does not transfer any knowledge to the other firm. Table 1.2 can hence be classified to account for the direction of knowledge flows.

Table 1.4 shows this classification. Just as the direction of the flow depends on the tube, so does the type of knowledge (tacit, implicit, public or private) that an transfer. For instance a social link, which is a discussion between employees can never allow for a transfer of physical capital, it allows information on problem solving to flow. An IP transfer is a transfer for codified knowledge only. Tacit knowledge needs repeated face-to-face interactions to be exchanged (Von Hippel, 1987). We hence have to understand which type of knowledge can be transferred through which channel to judge the potential importance of the transfer on the innovation process of the firm. Let us take a look at the channels and the objects that they might carry.

Tube	Bilateral	Unilateral	Both
License		■	
Joint Venture	■		
Alliance		■	
Social			■
Spin-Out		■	
Externalities	■		
Buyout		■	
Supplier	■		
OEM	■		
IP transfer		■	
R&D contract			■
Employee mobility		■	

Table 1.4: *Direction of knowledge flows per tube*

Table 1.3 allows us to see that some channels only allow for a specific type of knowledge to flow. For example in the case of social interactions only solutions to problems may flow. We have to emphasize here that an interaction between agents can imply different channels at the same time. A social link might exist at the same time as a licensing agreement, or the creation of a licensing agreement might lead to a social link. This does not change the information in our table. The different types of links are separate and transfer knowledge in their own right.

Channels involving research institutions

Many of the channels involving firms can also involve research institutions. The channel itself is not affected but the motivations are. Research institutions are typically at the pinacle of scientific knowlegde. A collaboration between a firm and a research institution is motivated, from the firm side, by a need for fundamental research and the expectation of radical innovation (Tödtling et al., 2009).

The RIs objective is different since it is less market oriented. A RI is motivated by the need for funding. Reputation is therefore important, the better the reputation the more funding a RI will be able to gather. Finding applications for the technologies that are developed and forming students to use them is hence of paramount importance. The reputation of a RI depends upon the quality and usefulness of the research conducted, a point where firms can play an important role. After all, firms have expert knowledge on market trends. In a

collaboration network, RIs have a central role in the sense that they add new knowledge to the knowledge base of the network ensuring that technological diversity does not decrease to a level where innovations would only be incremental.

Conclusion Networks are build up from linked agents. Even thought in the vast majority of analyses and models all these links are considered identical, in reality they are all different.
The diversity of channels is important because they each carry different types of knowledge that influence the R&D process of the firm in a different manner. Indeed, some of the links impose no particular restriction to the knowledge that may flow between collaborators others allow only specific types of knowledge to flow. This is due to fact that tacit knowledge needs time and regular interactions to be transferred. Channels that do not have this characteristic hence restrict the flow of tacit knowledge while it has an important impact on the productivity of the firm.

We can hence summarize channel's characteristics by 2 factors, the breadth and the length of the channel. The breadth would define the amount of knowledge that a firm is willing to exchange (which it reveals by offering a contract of a certain type) and the length defines the amount of knowledge that actually flows through the channel. In such a framework codified knowledge only needs a narrow and short channel (social link suffices) whereas mastering the large hadron collider needs a very broad and long channel for employees to learn a technology.

The diversity in channels also teaches us that several networks might exist at the same time. We refer here mainly to the social and the formal network. Formal interactions imply social interactions. The exchange of information through the social network is however different from that that transfers in the formal network. Information relative to trust and reputation flows through a social network that will have a different structure than the formal network of cooperations.
In any case the diversity of both channels and agents in the network defines the quality of the knowledge flow in the network and as a result defines the performance of the network as a whole.

1.5 Network efficiency

The strength of the network resides in its ability to evolve over time, getting rid of bad elements and innovate continuously by sending relevant knowledge to the firms composing it. The efficiency of a network can be understood in two ways: the speed and quality of the diffusion of knowledge, and the optimization of social surplus generated by collaborations through market interaction. In this section we will review elements that impact the efficient diffusion of knowledge through the network as well as more market oriented measures of efficiency (profit, utility, return on assets). Since the structure of the network has a vital role to play, one would wish to identify efficient structures for networks. Since efficiency is very difficult to measure empirically, theoretical models are often used to assess efficiency.

1.5.1 Equilibrium structures

The previous sections have discussed factors that impact the transfer of knowledge in a network. The structure of the network itself also impacts the efficiency of knowledge flows. In a sparse network structure knowledge needs more time to diffuse while a more dense structure allows for faster diffusion. Models of knowledge diffusion understand efficiency as the speed and quality of the transfer of knowledge through a network. The previously discussed elements hint however to the idea that diffusion highly depends upon the abilities of the agents transferring the knowledge. In order to have a better understanding of the effects of the structure of the network on network efficiency one is required to include other elements such as profit (König et al., 2012; Jackson and Wolinsky, 1996), utility (Jackson, 2003) or R&D expenses (Goyal and Moraga-Gonzalez, 2001). The latter calls for models that include market interactions between firms. Firms evolve on a market and face the same demand. They instigate collaborations that result in a reduction of their production cost (through knowledge flows), making them more competitive on the market. Firms continue to add links as long as the marginal benefit from a link exceeds that of the marginal cost. In order for a link to exist both firms need to accept to maintain it. If one of the firms decides the link is not beneficial it can unilaterally cut it. When no firm wishes to sever a link or add a link, the network is called stable. This specific concept of stability is called "pairwise" stability (Herings et al., 2014; Jackson and Wolinsky, 1996; Bala and Goyal, 2000; Jackson and Yariv, 2007). Different papers identify different stable

structures (stars, complete graphs, empty graphs). The question is then wether any of the identified stable structures are efficient (according to a particular criteria). When it comes to the diffusion of knowledge, the small world structure has been identified as the most efficient (Verspagen and Duysters, 2004). When one includes the industrial dimension, i.e how firms use the knowledge they receive in order to make a profit, results are much more ambiguous. König et al. (2012) show that the efficient network structures (as measure by the total profit of the network) are not necessarily the stable network structures. It turns out then, that firms are not able to organize themselves in manner that maximizes social surplus, due to their myopic, short-term, profit maximizing vision.
It turns out then that the role played by the structure of the network is more than ambiguous. Maybe, it is not so much the structure that is important but more the agents one is connected to. Network efficiency might simply be a question of efficiency of partner selection. A small world, identified as efficient in terms of knowledge flow, might be much less efficient if certain firms with strategic positions do not transfer knowledge efficiently.
We would hence be interested in models that allow for a more strategic partner selection mechanisms coupled with more clearly defined knowledge transfer mechanism. These considerations will however only show their fundamental importance if heterogeneity is introduced in the model. As we have seen previously the value of the network lies in the diversity of agents and the diversity of channels that connect them, allowing them to innovate.

The literature shows us that the structure of the network is highly dependent upon the partner selection mechanisms as well as sectorial aspects. The latter point is influenced by the presence of different types of agents present in the network. When public research institutions are present, they usually take a central position in the network impacting the average distance of the network. Some sectors rely more on these research institutions than others. High technology sectors will for example rely more on the presence of universities than would the manufacturing sector. The differences in the network should hence be visible through sectoral differences in which the networks evolve. The characteristics of the sector of activity will define the types of agents present and even the channels that are created between the agents. Some sectors will rely more on joint ventures because of high competition between firms, others will rely more on long term cooperations with suppliers and universities.

1.5.2 Obstacles and accelerators of efficient knowledge diffusion

Absorption capacity

When exposed to new knowledge a firm will want to absorb all knowledge that is useful to it. Even though the quality of the information it has access to is high, it might only be able to learn a small fraction of it. This inefficiency in the transfer stems from a low absorption capacity; the ability of the firm to learn. This means that it is completely dependent upon the technology level of the firm. The more technologies a firms already masters the easier it is for the firm to learn a new technology.

Tsai (2001) shows that the absorptive capacity of the firm is directly related to the business performance and innovation of the inter-firm units in a study covering a petrochemical company and a food manufacturing company. Østergaard (2009), for example, shows that the absorption capacity increases the probability of knowledge acquisition, this is reinforced by results from Giuliania and Bella (2005). Giuliani et al. find that firms with a higher absorption capacity are more likely to create links. The more links a firm has, the more (potential) access it has to knowledge. This capacity not only increases the innovative performance of the firm, it also reinforces the competitive advantage of the firm (Chen et al., 2009).

When studying the efficiency of innovation networks the absorption capacity is often casted aside because of the heterogeneity it introduces when researches want to focus on a specific aspect of innovation networks. Using an agent based model in which firms create alliances based on absorptive capacity rather than social capital, Egbetokun and Savin (2013) show that the resulting networks have a similar structure as the networks that result from social capital considerations. They also found that there was a positive correlation between the position of the firm in the network and the absorption capacity of the firm. In order to use all the available information in the network (either through tubes or by spillovers) firms need a high absorption capacity (Camisón and Forés, 2011).

Firms with a low absorptive capacity in a central position in the network can hence stop the diffusion of ideas and technologies, or slow it down significantly as happens with Chinese whispers. A network is hence only truly efficient if the agents composing it are indeed able

to learn from each-other.

Cognitive distance convergence

As I discussed before, knowledge exchange is one of most important motivations for the emergence and efficiency of innovation networks. In order to be able to learn from each other firms need a low to average technological overlap or short cognitive distance. When firms collaborate, they exchange knowledge which increases their overlap. The more they exchange the less unique knowledge they have, decreasing the diversity of knowledge. When this diversity becomes too low or disappears, the return to innovation then decreases. Wuyts et al. (2005) show that the technological overlap is a decreasing function of the frequency of cooperation. This highlights once again the importance of the presence of RIs in the network that allow firms to learn new technologies and diversify their abilities and, through recombination, find new technologies to avoid the convergence to a network in which all firms master the same technologies and innovation eventually dies.
Granovetter shows this in his seminal work in 1973 Granovetter (1973) in which he argues that the strength of a tie is positively correlated with the time spend between actors, showing how socially close they are. Granovetter applies the idea in sociology but the idea can be directly translated to the analysis of the firm. The higher the strength of a tie in this case the more efficient the transfer, at the same time the repeated interaction also means that firms have less to learn from each other and will eventually master the same technologies as we pointed out before. The weak ties hence play an important role since they bring this diversity to the neighborhood of the firms. The channels we enumerated previously hence have to be taken into account in a dynamic vision as well. One alliance at time t cannot be compared to a continuous alliance over time, the possibility for exchange is high but the risk of a lock-in is high. Lock-ins can occur not only at a technological level but also on a social level, firms might be afraid to select a partner.
This shows that if firms are too embedded in their network, their connection to the rest of the network is restricted. New ideas and technologies will take time to reach the firm because knowledge needs to travel to a dense network.This phenomenon, called overembeddedness, is a risk for firms in a network configuration.

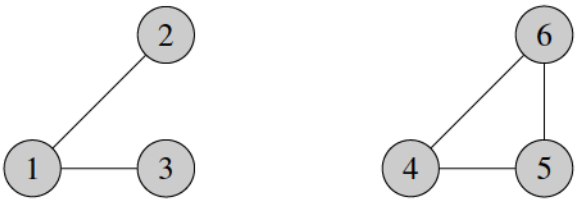


Figure 1.1: *Brokerage and Closure illustration*

The structure of the whole network and that of individual firms, hence plays a vital role in the diffusion process and thus the efficiency of the firms evolving inside the network. To complete the picture of the efficiency of innovation networks we will now see how the structure of the network influences the performance of innovating firms.

The theory of strong and weak ties and the importance of structural holes

The study of the structure of the network structure will allow us to assess the efficiency of information flow and hence the efficiency of firms inside the network.

We have stressed the importance of diversity of knowledge in a network. Some structures allow for this diversity to exist others do not. For instance suppose the networks in figure 1.1.

Even though the structure is simple the difference between these two structures highlights a vital point in terms of efficiency in cooperations. The structure of network 1 is very advantageous in terms of diversity of knowledge. Since A and C do not communicate they have different technologies and hence will improve the potential result of the R&D procedure. Firms B is in a very important position, B allows knowledge to flow from A to C indirectly.Without B there would only be two separate clusters in the network. This very advantageous position is referred to as a brokerage position or the bridging of a structural hole (Burt, 2004). A firm in a brokerage position connects different components of a network together. Not only is this position advantageous for the firm in terms of network importance, it also allows the firm to have first hand knowledge of new technologies or ideas. In a purely social context Granovetter (1973) shows that such a bridging link can only be a weak link. A weak link is a connection between agents that do not interact on a regular basis, they are mere acquaintances. Consider figure 1.2.

Agent A is in a brokerage position, if A had strong ties with C and B (socially close, they know each other well and interact often) then there is at least a weak link between C and B. The presence of this link results in the disappearance of the brokerage position of

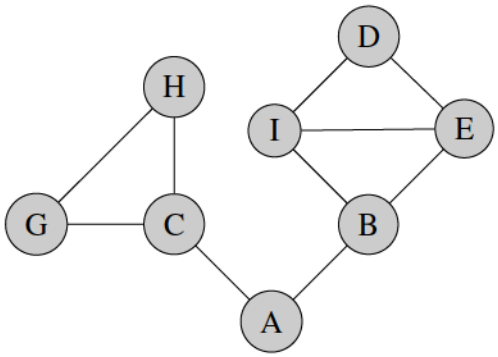


Figure 1.2: *Structural hole illustration*

firm A. Hence if a firm is in a brokerage position it can only have weak links. The theory of weak links and brokerage focuses on the positive impact of diversity. Hargadon (2002) show for example how brokerage positions held by consultants allow them to introduce new techniques in different sectors. At the antipode of this theory we find the theory of closure and strong links. In network 2 we observe closure, or the absence of structural holes. Granovetter suggests that closure allows for firms to better cooperate through the creation of norms between the agents. This idea can be translated to the analysis of the firm. When firms cooperate often and with the same firms or people there will be convergence in their methods and routines. This advantage of redundancy is counterbalanced by a positive influence of brokerage though the advantage of diversity in knowledge.
Empirically both theories are supported, however, one has to be careful when using the terminology weak and strong ties. Throughout the literature on networks and the literature on knowledge exchange a weak tie does not follow the definition of Granovetter. Notably we find the definition of Hansen (1999) for whom a weak link between firms is a unidirectional link between firms while a strong link is a bidirectional link. With this definition he found that weak links speed up projects in which knowledge has a low complexity while strong links sped up projects in which knowledge was complex.

Empirical evidence on the importance of structural holes does however exist in the literature on inter-firm networks. Ahuja (2000) found a negative influence of diversification on firm performance (measured by patent count) while Cohen and Levin (1989) found a positive influence. An analysis adjusting for the strength of the ties could here be of use. It is possible that the reason why redundancy in Ahuja's case does not have a positive effect comes simply from the fact that firms have weak ties and hence norms have not

emerged yet. Another explanation might reside in the fact that one has to distinguish between social links and formal links between organizations, as I have shown, the transfers are not the same when one considers a social link or an alliance and hence the influence on the performance of the firm will be different.
Shan et al. (1994) show that in the case of start-ups social capital is a better predictor of co-operation, they theorize that structural holes are efficient in the case of market transactions since there is no need for extensive cooperation over time. In contrast, by analyzing on the level of the individual, Burt (2004) find by the means of interviews with managers that there is a correlation between the brokerage position of managers and their productivity in terms of coming up with good ideas. Promotions and compensation were disproportionally given to managers that found themselves in a brokerage position in the network.

Overall the theory on strong and weak ties teaches us the importance of redundancy in the innovation process. It enhances the ability to cooperate by the creation of norms between firms while at the same time reducing the efficiency of new innovations by the reduction of diversity. This then shows the importance of structural holes. When we try to connect this theory to standard network theory (in the mathematical sense) we find that the problems surrounding redundancy are similar to the concept of clustering. The clustering of a graph is indeed the propensity of a graph to have triangles. The higher the clustering of a graph the more triangles, the higher the redundancy. The clustering coefficient of a graph can hence be interpreted in terms of redundancy or in terms of norm emergence.

The influence of structural holes on the performance of the network as a whole is however not clear. However, brokerage positions have shown to be beneficial for agents. They appear to be able to exploit their favorable position.

Overembeddedness

The existence of structural holes shows us the risk of their absence and the negative effects that this absence might have on firms' performance. Firms might become too embedded in their network which puts them in a position far away from new sources of information.

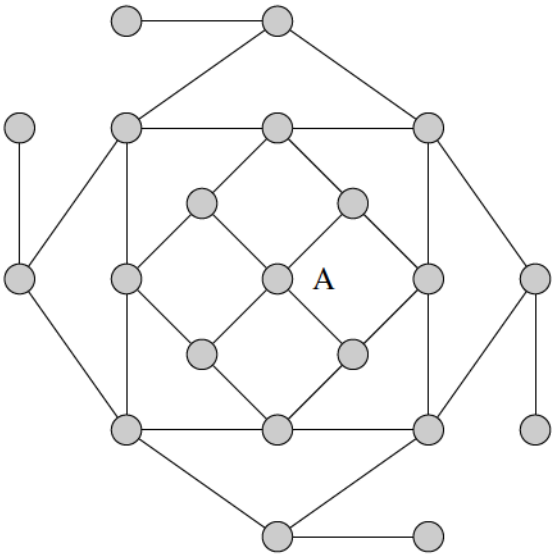


Figure 1.3: *Embeddedness illustration*

Consider for instance figure 1.3. The network is somewhat stereotypical, but is serves as a good example. Firm A is in a very central position in the network, the disadvantage of this position is that it is 5 links away from any new knowledge, moreover that knowledge will be recombined by knowledge it has in common with other firms reducing it efficiency in terms of diversity. A firm in such a position can be caught in a network lock-in meaning that it will not be able to find new partners to work with and will hence always work with the same firms which will result in a convergence of technologies used and reduce the efficiency of the innovation process of the firm significantly.

The question that is then raised is what structure is more efficient for knowledge transfer ? Different authors seek an answer either with a theoretical model or with a empirical analysis.

Empirically Innovation networks seem to be locally clustered (Geenhuizen, 2008; van der Valk et al., 2011), Asymmetric and sparse. Different canonical network structure have also been identified empirically: Small worlds and scale-free networks. Based on these empirical observations, models of knowledge diffusion have tested the efficiency of both the network characteristics and network structure. Small worlds are identified as the most efficient structure, both empirically and theoretically (Cowan and Jonard, 2007; Verspagen and Duysters, 2004; Gulati et al., 2012; Alghamdi et al., 2012). This observation can be explained by the fact that a small world structure is defined by a low average distance and

a high clustering coefficient. The combination of these two characteristics leads to a fast diffusion throughout the network.
The latter is however contrasted by a paper using a more complex method for the identification of optimal network structure. Using a genetic algorithm Carayol et al. (2008) find intermediary network structures to be more efficient.
When it comes to theoretical models studying the inception of networks, other structures are found to be optimal. Based on a cost-benefit analysis to link creation the first models (Goyal and Moraga-Gonzalez, 2001; Jackson and Wolinsky, 1996; Jackson, 2005) resulted in an equilibrium analysis with three possible solutions; empty network, star network or complete network depending on the cost of a link.
In most of these models firms select their partners at random. In reality however, the selection procedure can include different and more complex factors.

Partner selection

The structure of empirical networks are the result of strategic decisions made by firms and research institutions. The motivations of a firm or RI to collaborate with a specific firm or RI can possibly be explained by different factors. I exclude here projects that are financed by a government agency. In the latter case some agents might be included purely for reasons that are included in the contract and hence not the result of strategic decisions on the part of firms or RIs.

The most notable factors are:

- Social proximity (includes trust): starting a cooperation with a firm with whom one has not yet collaborated involves a high risk factor. First, the quality of the potential partner might not meet expectations. Second, on a social level, the employees that will be in contact with each-other might not get along well. And finally other firms might have different routines, converging to commun understanding might take time. For these reasons firms can prefer working with historic collaborators, reducing the risk of a failure.

- Technological proximity (Cognitive proximity): innovations are based on the recombination of technologies. As we will show in chapters 4 and 5, the technological distance between firms place a vital role in the decision to collaborate. When this

distance is too far apart firms will not be able to understand each-other and hence the recombination of technologies is inefficient. The relation between technological proximity and probability to collaborate has an inverted U-shape.

- Geographical proximity: In addition to being a possible condition in research contracts, geographical proximity has advantages for cultural and cost reasons. International collaborations can prove to be difficult because of cultural differences including work ethics and different practices. In addition, having one's collaborators close by allows for more frequent interactions and improve social proximity.

- Fitness / performance: When the previous factors are not leading to a decision, an agent would look at the performance of a potential collaborator. Technologically lagging firms can present a risk for the outcome of the R&D process. In the case of bankruptcy the R&D efforts would be lost.

The decision making process of a firm takes into account several of these factors. Which factor overrules another might depend simply on the project (and its aim) at hand. The efficiency with which knowledge flows through the network does depend heavily on the selected partners. When collaborators are chosen poorly, absorptive capacity as well as sending capacity might act as an obstacle to the flow of knowledge. Repeating collaborations with the same firms increases trust and the convergence of working methods, however this has the downside of reducing knowledge diversity for the network.
Despite the fact that firms are mainly not aware of their position inside innovation networks (or even social networks for that matter), their collaboration decisions alter the structure of the network in which they evolve. This is one of the manners in which the firms can themselves impact the network. Networks can have a stable structure if social interactions have a high impact on partner choice. The inherent risk factor can push firms to continue collaborating with historic partners rather than new ones (Gulati, 1995) . Trust hence place a vital role in the determination of the structure of the network (Ahlström-Söderling, 2003; Schrader, 1991). When interactions are repeated the cost of link maintenance can decrease while increasing the level of trust between firms. Human and Provan (1997) show that the structure of a network will not be the same when trust (or historic partners) or absent. The duration of these interaction does however not play an important role in terms of knowledge transfer Schrader (1991).

1.6 Conclusion

Networks emerge by strategic decisions of firms. Cooperations influence the innovative ability of the firms at every stage of the R&D process by the transfer of knowledge and information. Different channels allow for the transfer of different types of knowledge and information. The transfer of reputation allows for the selection of the optimal partner. The latter allows for increased knowledge flow and efficiency in terms of collaboration. In this sense the firm shapes the network. Efficient networks, measured theoretically, can take a large variety of forms and depend upon the definition of efficiency. In terms of knowledge diffusion the small world structure appears to be most efficient while in terms of profit the structure is ambiguous. These results are however based on models that do not fully take into account all the accelerators and obstacles to knowledge diffusion. It could then be possible that there is no such thing as an efficient network structure. The efficiency of a network is largely dependent upon the specific position of certain firms inside the network. Identical structures with differently positioned firms can behave differently. Of course, it might be that firms with better positions reached their position because they were more efficient to start with. The efficiency of the network and its ability to survive hence depends upon the ability of the firms to learn from one another and on their ability to make optimal partner choices (by avoiding social lock-ins) to avoid the diversity of technology to run out. Context is of paramount importance when studying a network. The vast majority of network analyses oversimplify the complexity of networks and hence are unable to extract all the relevant information included in a network. A network is more than a simple aggregation of bilateral cooperations and this should be reflected in any conducted analysis. The firms, as a part of the network, are influenced by it, a symbiosis exists between the network and the agents that compose it.

The analysis of this symbiosis starts with an understanding of the dynamics of link creation. The decisions that brings a firm to collaborate with another specific firm shapes the network and informs us about the manner in which innovation is achieved. The decisions made by firms result in a particular position inside the network that could favor or hinder the performance of the firm. The next part of this thesis is hence dedicated to the analysis of

the dynamics of the structure of innovation networks and how firms perform according to their position inside the network. The question of the overall performance of the network will be treated in the final chapter of the thesis with a theoretical model.

Chapter 2

Introduction to network modeling using Exponential Random Graph Models (ERGM)

"Come on, Rory! It isn't rocket science, it's just quantum physics! –The eleventh Doctor"

In chapters 4 and 5 a method of network analysis is used that is not yet widely used. The purpose of this chapter is to explain the theory behind this method before applying the method in the next chapters.

Introduction

Networks are representations of relational data. Nodes represent entities while the links connecting them represent any form of interaction or connection between

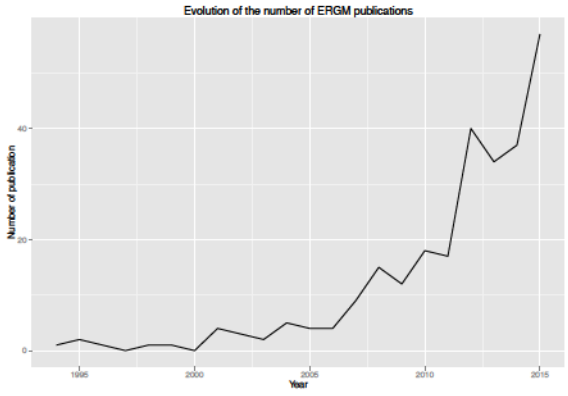


Figure 2.1: *Evolution of the number of publications involving ERG models for all disciplines (statistics included) (source: Scopus)*

the entities. A large diversity in the types of networks exists ranging from networks of social contacts between individuals to inventor networks, collaboration networks, financial networks and so on. The trouble with networks, especially when represented by a graph, is that it looks like a large heap of lines, it resembles pure chaos. However, interactions between individuals, firms or banks rarely appear at random. The motivations for link creation cannot be observed directly from a graph, nor are they clear from glimpsing at a database containing relational data. In order to identify the motivations for entities to create links and identify the global network structure, a more in-depth analysis is required. Econometric analysis could shed more light on the motivations behind an observed link through logistic regressions. The probability of a link could be explained by a number of variables. There is one important limitation to this method. Due to the hypothesis of independence of the observations the probability of a link between two nodes can never be explained by the presence of another link inside the network. It is feasible that a link between two nodes exists only because of the presence of other links in the network. Take for instance the idea that John and Mary are connected solely because they have a common contact: Paul. ERGM models are modified logistic regressions that allow for the probability of a link to depend upon the presence of other links inside the network (amongst other variables of course). An ERGM identifies the probability distribution of a network so that it can generate large samples of networks. The samples are then used to infer on the odds of a particular link inside a network. Applications for this method are numerous in many fields of research as shown by the increasing trend in the number of publications using ERGM models (see figure 2.1). In economics the number of published papers appears to be relatively low when compared to the other social sciences. Only 19 published papers could be found in the Scopus database (and even less in the web of science database). The topics are however quite diverse: Link between money and inflation (Tang et al. (2015); Özdemir and Saygili (2009); Czudaj (2011); Belke and Polleit (2006); Price and Nasim (1998)) , knowledge sharing in organizations (Caimo and Lomi (2015)), GDP targeting (Belongia and Ireland (2014)), alliance networks (Cranmer et al. (2012); Lomi and Pallotti (2012); Lomi and Fonti (2012); Broekel and Hartog (2013)), geographic proximity (Ter Wal (2013)).

The growing interest, and development of a theory of economic networks, provides a fertile ground for the use of ERGM models from the geography of innovation to venture capital

investments. The aim of this chapter is to provide an overview of the basic statistical theory behind ERGM models, which will be dealt with in the first section. Section 2 discusses the concept of dependence and the explanatory variables that can be included in the models. Section 3 discusses estimation methods while section 4 provides the R scripts and the interpretation of an example using data for the French aerospace sector alliance network.

2.1 Theory

2.1.1 The canonical form of ERGM models

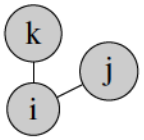


Figure 2.2: *Network G*

The aim of an ERGM is to identify the processes that influence link creation. The researcher includes variables in the model that are hypothesized to explain the observed network, the ERGM will provide information relative to the statistical significance of the included variable much like a standard linear regression.

It is useful at this point to explain that sub-structures of a network can (and are predominantly) used as explanatory variables. Substructures are small graphs contained inside the network. Examples can be found in figure 2.28. The presence of some of these structures reflects certain link creation phenomena. A random network, i.e a network in which a link are created at random, show a low number of triangles. A triangle is an interconnection of three nodes, the smallest possible complete subgraph. The presence of triangles in an empiric network bares witness that there is process that generates triangles that is not the result of random link creation, e.g a tendency to create link between common friends. A network with a small number of large stars and a large number of small stars can be the results having a small number of very popular nodes. This is found in citations networks as well as lexicographical networks. Including sub-structures allows the modeling of certain processes as would any other variable.

In an ERGM we can find two types of explanatory variables: structural and node or edge-level variables. The latter come from other data sources and can be for example age,

size of a firm, proximity, gender and so forth.

2.1.2 The odds of a link

With this in mind the probability that one would observe a link between any two nodes i and j in a given network is proportional to a given set of explanatory variables:

$$p(G_{ij} = 1) = \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \ldots + \theta_n \cdot X_n \tag{2.1}$$

We note G a graph, ij the focal nodes. $G_{ij} = 1$ means that a link exists between nodes i and j in graph G, $G_{ij} = 0$ implies the absence of a link between nodes i and j[1]. θ is a vector of parameters and X a vector containing the variables.
This equation gives us the probability of a *single link* in graph G. Since nothing guarantees the probability to stay within $[0, 1]$ the equations needs to be somewhat modified. We start by transforming the probability into an odds ratio:

$$odds(G_{ij} = 1) = \frac{p(G_{ij} = 1)}{1 - p(G_{ij} = 1)} = \frac{p(G_{ij} = 1)}{p(G_{ij} = 0)} \tag{2.2}$$

Now, in equation 2.2 we notice that the odds of a link tend towards zero when the probability of a link tends towards one, while tending towards $-\infty$ when the probability tends towards zero. The final modification ensures that the distribution stays between the bounds of zero and one, this is accomplished using the natural logarithm:

$$log(odds(G_{ij} = 1)) = logit(p(G_{ij} = 1)) = log\left(\frac{p(G_{ij} = 1)}{p(G_{ij} = 0)}\right) \tag{2.3}$$

The probability is now bounded by zero and one. If we suppose that the probability of a link is explained by a vector of n variables accompanied by their respective parameters $(\theta_1 \ldots \theta_n)$ then we can write:

[1] the values 0 and 1 refer to values found in an adjacency matrix, 1 indication the presence of a link, 0 the absence

$$logit(p(G_{ij} = 1)) = \theta \cdot X \tag{2.4}$$

So in our example we have:

$$\begin{aligned} logit(p(G_{ij} = 1)) &= \theta_1 \cdot X_1 + \ldots \theta_n \cdot X_n & (2.5) \\ p(G_{ij} = 1) &= exp\{\theta_1 \cdot X_1 + \ldots \theta_n \cdot X_n\} & (2.6) \end{aligned}$$

The logit gives the marginal probability of a tie between nodes i and j. However, if we were to compute the probability of a link between i and j the probability for i and k the results would be independent. A logistic regression works under a hypothesis of independence of observations. In the case of networks observations are not independent. For instance, common friends tend to connect more in social networks, common collaborators have a higher tendency towards collaboration. A model that aims at explaining a network structure should be able to include these tie formation processes.

We hence modify the initial equations to include the network structure as observed before the link. This modification is introduced by Strauss and Ikeda (1990). We note G^c_{ij} the network without link ij:

$$odds(G_{ij} = 1) = \frac{p(G_{ij} = 1|G^c_{ij})}{1 - p(G_{ij} = 1|G^c_{ij})} = \frac{p(G_{ij} = 1|G^c_{ij})}{p(G_{ij} = 0|G^c_{ij})} \tag{2.7}$$

In equation 2.7 the odds of a link between nodes i and j now depends on the structure of the network before a link between i and j is created (noted by $|G^c_{ij}$). The probabilities are now conditional.

We discussed previously that some of the variables in the model can be subgraphs. The manner in which these are included in the model is simply by the count of these substructures. In other words, the value of the variable triangles is the number of triangles in the network. The same is true for stars, circuits and shared partners. This has as a consequence that the counts of these variables are not the same when a link between two nodes is present or absent. For instance the number of edges changes by one. This means that we need to differentiate between the value of the variables when a link is present and

when it is absent. We hence note the vector of variables $v(G_{ij}^{+})$ when a link between i and j is added (hence the " + ") and $v(G_{ij}^{-})$ when the link is absent. By including this differentiation we can rewrite equation 2.7 using the result in equation 2.6:

$$odds(G_{ij} = 1) = \frac{p(G_{ij} = 1|G_{ij}^{c})}{p(G_{ij} = 0|G_{ij}^{c})} = \frac{exp\{\theta' \cdot v(G_{ij}^{+})\}}{exp\{\theta' \cdot v(G_{ij}^{-})\}} \tag{2.8}$$

Where $v(G_{ij}^{+}$ represents the vector of variables in the network with the link between i and j present and $v(G_{ij}^{-}$ the vector of variables with no link between i and j.
With some basic algebra we can develop the previous equation a bit further:

$$\frac{exp\{\theta' \cdot v(G_{ij}^{+})\}}{exp\{\theta' \cdot v(G_{ij}^{-})\}} = exp\{\theta' \cdot v(G_{ij}^{+})\} \cdot exp\{-\theta' \cdot v(G_{ij}^{-})\} \tag{2.9}$$

$$= exp\{\theta'(v(G_{ij}^{+} - v(G_{ij}^{-})\} \tag{2.10}$$

When developing the vector of variables we have:

$$= exp\{\theta_1' \cdot (v_1(G_{ij}^{+}) - v_1(G_{ij}^{-})) + \ldots + \theta_n' \cdot (v_n(G_{ij}^{+}) - v_n(G_{ij}^{-}))\} \tag{2.11}$$

Equation 2.11 shows that each parameter of the model is associated not with the counts of sub-structure but with the difference in counts. The difference from having an extra link, and the absence of said link. In essence $(v_1(G_{ij}^{+}) - v_1(G_{ij}^{-}))$ represents the variation in the number of counts of network statistic 1 that result from the additional link. The variables are hence referred to as "change statistics". In order to remove the exponential from the right hand side of the equation we apply the logarithm:

$$log\left(\frac{exp\{\theta \cdot v(G_{ij}^{+})\}}{exp\{\theta \cdot v(G_{ij}^{-})\}}\right) = \theta_1' \cdot (v_1(G_{ij}^{+}) - v_1(G_{ij}^{-})) + \ldots + \theta_n' \cdot (v_n(G_{ij}^{+}) - v_n(G_{ij}^{-})) \tag{2.12}$$

So we can rewrite equation 2.12 noting $v_1(\Delta_1 G_{ij})$ the change statistic for a link

between i and j for variable 1 as follows:

$$log\left(\frac{exp\{\theta \cdot v(G_{ij}^{+})\}}{exp\{\theta \cdot v(G_{ij}^{-})\}}\right) = \theta_1' \cdot v_1(\Delta_1 G_{ij}) + \ldots + \theta_n' \cdot v_n(\Delta_n G_{ij}) \tag{2.13}$$

Each variable now accounts for the change in counts of network statistics. It is important to remind us that equation 2.13 accounts for the odds of one edge in the network while we are interested in the probability for the whole network. Following (Besag, 1972) we can invoke here the Hammersley-Clifford theorem. Since this theorem is based on concepts out of the reach and the purpose of this document we will not detail the theorem. For a detailed explanation please refer to (Hammersley and Clifford, 1971).
The theorem states that the probability of a network can be defined solely by the counts of subgraphs. This is important because it tells us that all we have to do is identify the correct subgraphs to ensure that a model of the network structure can be found. The more accurate the subgraphs to more reliable the inference of additional covariates.

2.1.3 The probability distribution of a network

The Hammersley-Clifford theorem states that the probability of a graph can be identified solely by counts of subgraphs. As such, we know that the probability is proportional to these variables. Since we have an observed network that we wish to replicate we look for the probability that the network generated by the model (X) is identical to the observed network (x). The logarithm is applied to bound the probability:

$$log(p(X = x)) \propto \theta \cdot v(G) \tag{2.14}$$

$$p(X = x) \propto exp\{\theta \cdot v(G)\} \tag{2.15}$$

The right-hand side of the equation now needs to be normalized in order to obtain a proper probability distribution. The normalization is not straightforward, indeed, in order to normalize the probability of a network one needs to normalize by all possible networks

with the same number of nodes:

$$p(X = x) = \frac{exp\{\theta \cdot v(G)\}}{\sum_{y \in Y} exp\{\theta \cdot v(G)\}} \tag{2.16}$$

With y a possible network structure in the set of all possible networks Y. The numerator is normalized by the sum of the parameters over all possible network structures.
Note that this number is large. For a network with n nodes the number of possible graphs is $2^{\frac{n(n-1)}{2}}$. So even for a graph with 10 nodes there are 35184372088832 possible graphs. The major problem to overcome with ERGMs is exactly this normalizing constant.

With some simple algebra we find a general form for this model. Using θ as a vector of parameters and $v(G)$ a vector of variables for network G:

$$p(X = x) = \frac{exp\{\theta \cdot v(G)\}}{exp\{log(\sum_{y \in Y} exp\{\theta \cdot v(G)\})\}} \tag{2.17}$$

$$p(X = x) = \frac{1}{\psi(\theta)} \cdot exp\{\theta \cdot v(G)\} = exp\{\theta \cdot v(G) - \psi(\theta)\} \tag{2.18}$$

Equation 2.18 is the most general and commonly used form of the model (Lusher et al., 2012). Equation 2.18 also gives the canonical form of an ERGM model. Since the density of the random variable (the network structure) has the particular form in equation 2.18 it is referred to as an exponential family. In additions, since the structures are represented by a random variable, they are random graphs.

Putting both elements together and this results in an Exponential Family Random Graph. Since ERGM is easier to pronounce than EFRGM, the models are referred to as ERGM[2].

The canonical form gives the equation we wish to estimate However, before we tackle the question of estimation we need to explore in more detail the variables that we would want to include in the model. We have stated previously that counts of subgraphs can be used as explanatory variables. The following section will explain which particular

[2]This type of model is also referred to as the P^* family of models (Anderson et al., 1999; Lusher et al., 2012).

subgraphs are to be included in a model.

2.2 The dependence assumption

The previous section has shown that ERGMs are capable of providing conditional probabilities for links. This dependence assumption is important because it allows the researcher to study different phenomena that rule the formation of networks. This section will show how the hypothesis of dependence of links is connected to the choice of subgraphs that may be included in an analysis.

2.2.1 The Markovian hypothesis

Links between nodes rarely appear at random, agents have specific reasons to connect with one entity rather than another. The motivations behind interactions are numerous and complex and have been subject to scrutiny from researchers in different strands of the social sciences. This research shows that people or firms with common acquaintances or collaborators will have a higher tendency to cooperate for example. This makes sense for two reasons, first, having common partners means common working practices which have a positive impact on collaboration. Second, when searching for collaborations firms tend to rely on referrals. A collaboration already in place allows firms to observe in detail the efficiency of other firms, referrals that result from cooperation should hence be trustworthy. In addition cooperators of a firms have a higher probability to be in contact with each other since they are more likely to meet during social or professional events.
The odds of a link depend on the neighborhood of the node and not on the entire rest of the graph. In more formal terms: two potential links X_{ij} and X_{kl} are in the same neighborhood if $\{i, j\} \cap \{k, l\} \neq 0$. For instance, the dotted lines in Figure 2.3 represent potential links. The odds of this link will depend upon the nodes f and d have in common, in this case node a. This shows that not all subgraphs are compatible with analyzing this phenomenon. Any substructure that cannot account for a common node should not be included.

Since it is hypothesized that only neighboring nodes impact the odds of a link, we seek node-level dependence. This level of dependence is also referred to as a nearest-neighbors level of dependence[3] or dyadic dependence (Harris, 2013).

[3]Nearest-neighbor systems have been studied by (Besag, 1972).

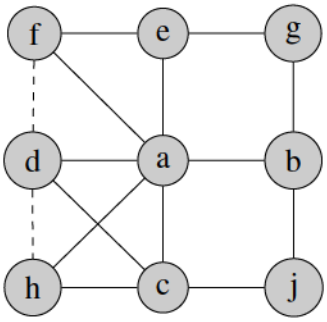


Figure 2.3: *Node level dependence illustration: the Markovian neighborhood*

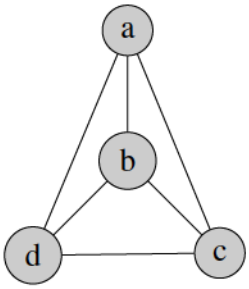


Figure 2.4: *Markov graph*

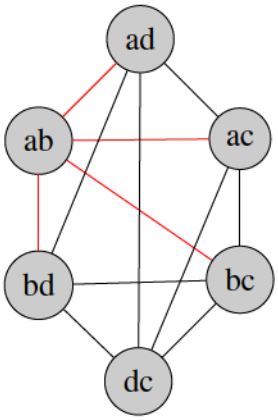


Figure 2.5: *Dependence graph*

For the purpose of the identification of relevant structures we need to find all structures that are subject to a Markov (or nearest-neighbor) level of dependence.

Suppose we have the social network depicted in figure 2.4. The graph shows social interactions between four agents, a, b, c and d. Markovian dependence suggests that a link between a and b depends on the connections between common nodes. The nodes a and b have in common are d and c. A link between a and b hence depends upon connections between $a - c$, $a - d$, $b - c$ and $b - d$.

When one identifies all the possible dependencies one can generate a dependence graph. The dependence graph for the complete Markov graph in our example can be found in figure 2.6. In red we find the dependence links for link $a - b$, it show the links on which $a - b$ depends.

From this graph one can identify the substructures that comply with Markovian dependence. All subgraphs in the dependence graph can be included in a model to add Markovian dependence. For example, the dotted line between ab and ad represents a 2-star centered on agent a (figure 2.7).

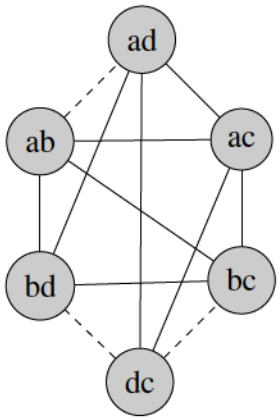


Figure 2.6: *Dependence graph and configuration identification*

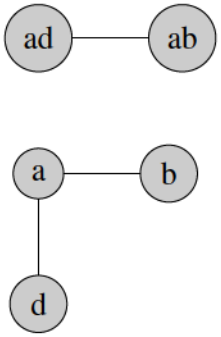


Figure 2.7: *2-star identification in the dependence graph*

Using the same method as for the 2-star one can also identify a triangle between the three agents on the links *bd*, *bc* and *dc*. The Markov model hence includes three configurations: edges, 2-stars and triangles. With the inclusion of these configurations the Markovian model takes the form:

In more complex graphs one could also identify 3-stars, 4-stars etc.

It should be obvious here that the number of distinct 2-stars is large and it is near impossible to add a parameter for each distinct 2-star in the dependence graph. To reduce the number of variables a hypothesis is made that each type of configuration has the same probability of appearance, this allows for the inclusion of one parameter per substructure. In the case of Markov dependence the ERGM model would have the following form:

$$p(x = X|\theta) = \frac{1}{\psi(\theta)} exp\{\theta_E \cdot v_E(x) + \theta_{S2} \cdot v_{S_2} + \ldots + \theta_{S_{n-1}} \cdot v_{S_{n-1}} + \theta_\Delta \cdot v_\Delta\} \quad (2.19)$$

Where θ_E is the parameter for the number of edges, θ_{S2} the parameter for the number of 2-stars and θ_Δ the parameter for the number of triangles. Note here that the model does not include simultaneously a 1-star and an edge parameter since they would be the same variable. With this model one is able to study if common nodes have a positive impact on the odds of link creation.

In addition the combination of the 2-star parameter and the triangle parameter account for triadic closure effects. In other words, are triangles created because 3 nodes are connected at the same time, or are triangles formed by the closing of a 2-star.

Of course, Markovian dependence is only one of the possible levels of dependence. One can imagine higher levels of dependence, or even any level of dependence to be empirically relevant. One could suggest that firms evolving on the periphery of a network to have a higher probability to connect with firms in the center of the network than between them. The previous model was hence extended by Wasserman and Pattison (1996) to allow for a general level of conditional dependence giving the researcher a total liberty in the theories to test. Whatever the level of dependence chosen, the dependence graph gives the substructure that may be included (Frank and Strauss, 1986).

Higher levels of dependence

It is possible to assume that the Markovian level of dependence is not adequate or does not capture the full complexity of mechanisms of link creation. Links can be dependent without there being a common node involved. For instance, consider a case in which people work on the same floor in a company. The probability of a social link does not depend upon a common node but simply on the fact that they are geographically close, belong to the same community or have common cultural aspects (White, 1992). In order to be able to model more complex aspects of social interactions and indeed even strategic interactions, one needs to be able to account for more structural aspects than stars and triangles (however potent in explanatory power these might be). The latter implies that the links are only dependent on each-other if nodes are part of a same neighborhood (neighborhood takes a broad definition here, it can be social, geographical or cultural). Due to the inclusion of general dependence the model is transformed to take the form:

$$p(x = X) = \frac{1}{\psi(\theta)} exp\{\sum_{A \in M} \lambda_A \cdot z_A(x)\} \tag{2.20}$$

Where A indicates the neighborhood as part of the ensemble M of all possible neighborhoods. The parameter λ_A will take a positive value when the probability of observing network x is increased. With the broad definition of "neighborhood" this model is able is almost limitless. The latter results in a problem, the model is too general.

We have seen in the previous subsection that the structures that can included in the ERGM model are defined by the dependency graph. In the case of a generalization of the dependency assumption, i.e all ties may be dependent upon all other ties, the dependency graph is a complete graph and all possible subgraphs can be considered as variables. This leaves a tremendous amount of parameters to be estimated.
Pattison and Robins (2002) and Carrington et al. (2005) offer two solutions to this problem. Their aim is to find a way to reduce the number of subgraphs to be included in the model. The only way to achieve this is to reduce the level of dependency from general to a more restricted level. A first step is to simply fix a level of dependency which will automatically switch all other parameters to 0. This means that once one defines a condition under which links are dependent upon each other a *setting* is defined. Defining a level of dependency can be simply supporting the hypothesis that links between firms depend upon a common region, or sector, size or any other group. s is a setting, s being a subset of the set of nodes M: $s \in M$. The restriction gives a new dependence graph which will contain a restricted number of subgraphs to include. All parameters for substructures that are not part of the dependency graph are equal to 0. Obviously, defining the settings oneself required extensive knowledge about the network at hand. The inclusion of these settings results in what Pattison and Robins (2002) refer to as *partial conditional dependence*.

Of course, one can also include other types of variables to a model, such as age of the firm, geographic location, amount of public funds received etc. These variables are referred to as node variates or node attributes The addition of these attributes is introduced by Robins et al. (2001). The idea here is that links depend upon the attributes of the nodes they are connecting. In other words the probability that two nodes are connected depends upon the attribute of the node. These models are also called *social selection models* and take the following form:

$$p(Y = y \mid X = x) = \frac{1}{\psi(\theta)} \cdot exp\{\sum_i \theta z(x) + \theta^a \cdot z^a(x, y)\} \tag{2.21}$$

Where the exponent a indicates parameters and structures relative to the inclusion of dyadic dependence for the purpose of social selection.

In the same paper Robins et al. also described a *social influence model* in which the attributes of the nodes are a result of the structure of the network (nodes take a particular attribute according to the nodes in the neighborhood for example). In other words, the probability that a node variables takes a particular value depends upon the structure of the network and the values of this (or indeed any other) node-level variable.
We hence need to add a node variables to the model. Suppose we note Y_i the value of a node variable for node i. This variable can be anything from a geographic region to the amount of public investment received by firms to the number of publications or patents. When this variable is included in the ERGM the model is written:

$$p(Y = y \mid X = x) = \frac{1}{\psi(\theta_i)} \cdot exp\{\sum_i \theta z(y, x)\} \tag{2.22}$$

Just as it is possible to put values on nodes it is also possible to values on dyads. The question then is to know if the value of the dyad increases the probability of nodes being connected. Think of situations where we would like to know if the amount of investment between firms is related to cooperation or if technological proximity between firms induces collaboration. Note here that the difference between dyadic covariates and actor attributes resides in the value on the link between two nodes. In the case of proximity is refers to the proximity of both firms, it is hence not a firm-level variable. The value only makes sense when we consider firms two-by-two.

All the extensions made to the ERGM framework allow researchers to answer a large variety of questions about social and economic processes. Many other extensions which are beyond the scope of this document, but worth noting, are multivariate relations in networks (Pattison and Wasserman, 1999) and dynamic networks in which new ties depend upon the structure of the network at time $t - 1$. It is also possible to model multivariate networks using ERGM. The idea is then that each link can exist in different matrixes, each corresponding to a different type of link (social, work, geography etc.). This extension allows researchers to study interplay between different networks and how each network is affected by the other networks.

Before looking at estimation methods for ERGM models one problem needs to be addressed: the degeneracy problem.

2.2.2 Solving degeneracy: Curved ERGMs

ERGM models are prone to degeneracy issues. When estimating the model the change statistics can behave in such a way that the large majority of the probability distribution is placed on either an empty or a full graph. As we will discuss in more detail a bit later, a simulation is performed to identify the probability distribution of a graph. This is done on a step by step basis, starting with an empty network and adding links one by one until a maximum likelihood is achieved. This probability distribution is a function of the change statistics and is thus impacted by the change in log-odds for an additional tie (for a given effect). In other words if an additional edge would create two new 2-stars, then the log-odds of observing that tie would increase by two multiplied by the parameter of the 2-star variable. A random graph model is considered stable if small changes in the parameter values result in small changes in the probabilistic structure of the model (Handcock et al., 2003). When a new edge is added to the graph this not only increases the number of edges but might also increase the number of other sub-configurations that might be included in the model. The parameters of the model control the number of sub-graphs of a particular type that are expected in the graph. For instance a 2-star might be transformed into a triangle by the addition of an edge which also adds two 2-stars. A 2-star can become at 3-star and so on. This cascading effect will result in the simulation procedure jumping over the MLE and converge to a full graph. Lusher et al. (2012) [chapter 6] show that, in the case of the Markov (or triad model), the number of expected triangles increases as the parameter for triangles increases. They highlight a phase transition for certain values of the parameter where the number of expected triangles increases more than exponentially. This transition explain that the probability density distribution has high weights either on a (close to) empty graph or on a complete graph[4]. This problem is increasingly present as the number of nodes increases. The larger the network the higher the number of substructures one can have.

In order to avoid the model to put too much weight on the full graphs, Snijders et al. (2006) propose to add several variables based on their concept of *partial conditional dependence*.

[4] In addition, the degeneracy of the model can result in problems with the estimation procedures.

The idea is to include a weighted degree distribution to the model, giving a high weight to low density while decreasing the weights as the degree increases. This reduces the impact of the high density variables responsible for the degeneracy of the initial model. Mathematically we can then write (using the notations of the initial paper):

$$u_{\alpha}^{(d)} = \sum_{k=0}^{n-1} e^{-\alpha k} d_k(y) \tag{2.23}$$

Where $d_k(y)$ is the number of nodes with degree k and α the parameter of the weights. This is referred to as the *geometrically weighted degree distribution*. The degree distribution can also be written as a function of the stars in the network. After all, a degree distribution is nothing more than a distribution of stars. Nodes with a degree of five are 5-stars, degree two are 2-stars and so forth. We can hence formulate the distribution as follows:

$$u_{\lambda}^{s} = S_2 - \frac{S_3}{\lambda} + \frac{S_4}{\lambda^2} - \ldots + (-1)^{n-2} \cdot \frac{S_{n-1}}{\lambda^{n-3}} = \sum_{k=2}^{n-1} (-1)^k \cdot \frac{S_k}{\lambda^{k-2}} \tag{2.24}$$

Where S_k is a the number of stars of degree k and $lambda$ the parameter. This method is referred to as *Aternating k-stars*. The difference between the geometrically weighted degree distribution and the K-stars is resides in the alternating signs. A large value of 3-stars is counterbalanced by a negative value for 4-stars due to the inverse sign of the parameter. The addition of the weights ensures that the change in change statistics stays small. Indeed Snijders et al. (2006) show that the change statistics can be written:

$$z_{ij} = -(1 - e^{-\alpha})(e^{-\alpha y_{i+}^{\tilde{}}} + e^{-\alpha y_{j+}^{\tilde{}}}) \tag{2.25}$$

Where $y_{i+}^{\tilde{}}$ is the density of firm i when the link between i and j is added. Equation 2.25 shows that the value of the change statistic is reduced by the factor $-(1 - e^{-\alpha})$. This factor hence ensures that the change statistics do not take too high values and result in a nested probability distribution. The inclusion of either the alternating k-stars or the geometrically weighted degrees transform the ERGM model into a Curved Exponential

Random Graph Model (Efron, 1975).

All that is left to do now is estimate the model.

2.3 Estimation

Estimation allows for the identification of the parameters that maximize the likelihood of a graph. Since we only have one observation (the observed graph) a set of graphs from the same distribution is required. The set of graphs that may be generated by this procedure should have the observed graph as a central element to ensure a valid sample.

2.3.1 Markov Chain Monte Carlo

In the first section we identified the general form of an ERGM model (see equation 2.18). The odds of a graph were normalized by the sum of the parameters of all possible graphs. This leaves us with a constant to estimate which is near impossible. A workaround has to be found for ERGMs to be useful. A first development by Besag (1975); Strauss and Ikeda (1990) was to estimate the model using pseudo-likelihood estimation. The properties of this method are however not clear (Snijders et al., 2006; Robins et al., 2007) we shall hence focus here on more recent methods that are better understood.

A method for estimating the parameters of ERGMs using a sample is developed by Anderson et al. (1999); Snijders (2002); Geyer and Thompson (1992). They estimate the model by Markov Chain Monte Carlo (MCMC) to find the maximum likelihood estimates (MLE). The idea is to extract a sample from a distribution that follows equation 2.18 asymptotically, not requiring the direct computation of the normalizing constant. Their paper points out that almost any maximum likelihood can be accomplished by a MCMC. A Markov chain is a sequence of random variables such that the value taken by the random variable only depends upon the value taken by the previous variable. We can hence consider a network in the form of an adjacency matrix in which each entry is a random variable. By switching the values of these variables to 0 or to 1 (adding or removing a link from the network) one can generate a sequence of graphs such that each graph only depends upon the previous graph. This would be a Markov chain. The hypothesis is then that if

the value at step t is drawn from the correct distribution than so will the value at step $t+1$. Unlike regular Monte-Carlo methods, the observations that are sampled are close to each-other since they vary by a single link. However, one would need a method for selecting which variable should change state in order to get closer to the MLE, this is done using the Metropolis-Hastings algorithm or the Gibbs sampler.

2.3.2 Metropolis-Hastings algorithm and the Gibbs sampler

The Metropolis-Hastings algorithm picks a variable at random and changes it's state. This results in either a new edge in the network or in the disappearance of an edge. The probability of the graph is then computed and only if the probability of the altered graph is higher than the previous one is the new graph retained for the next step. In other words the new graph is retained as long as the likelihood is increased:

$$min\{1, \frac{p_{\theta(x^*)}}{p_\theta(x^{m-1})}\} \tag{2.26}$$

This decision rule is called the *Hastings ratio*. The advantage of this ratio is that it does not include the normalizing constant $\psi(\theta)$.
Since the Markov chain starts at 0, a burn-in in needed to remove part of the chain to identify if the chain has converged or not (the burn-in can be parameterized in most software). The steps taken by the Metropolis-Hastings algorithm are quite small. These small steps are implemented in order to avoid overstepping the global optimum which can easily happen in the case of larger parameter spaces. Other methods allow for bigger steps and as such converge faster and need a lower burn-in. The risk of larger steps is however overstepping the global optimum and convergence towards other local optima. The Metropolis-Hastings algorithm may be slower than others but is more precise in it's estimation.

Some programs use the Gibbs sampler, which is a special case of the Metropolis-hastings algorithm (Hunter and Handcock, 2006). The difference between Gibbs and Metropolis-Hastings resides in the chosen step. In the case of the Gibbs sampler, the state of each element in the vector of parameters is chosen and updated conditionally on the state of the other parameters. This means that if this decision rule was implemented in the

Metropolis-Hastings algorithm the probability that the change is retained is equal to one. This makes the Gibbs sampler a relatively fast method. This sampling method is used by the different algorithms that are used to estimate ERGM models.
Both methods allow for the generation of a sample of graphs that can be used for inference. The sample of graphs is obtained by varying not the parameters but the variables of the model until it is centered around the observed graph. Now that a sample of graphs has been found we need to estimate the parameters of the model. Two of the most widely used estimation algorithms, the "Stepping" and "Robbins-Monro" algorithm will now be reviewed.

2.3.3 The "Stepping algorithm"

This method introduced by Hummel et al. (2012). It has the advantage of approaching the MLE directly while "Robbins-Monro" does not. ERGM models are indeed estimated using the maximum likelihood method. Starting from the canonical ERGM form we define the log likelihood function as:

$$L(\theta) = \theta \cdot v(G) - log(\psi(\theta)) \tag{2.27}$$

The problem here is the presence of the normalizing constant which cannot to computed. The improvements of this method over the previously one resides in the use of a log-normal approximation. The algorithm proposed here will converge towards the log-likelihood using a step-by-step method. The sampler used in with this estimation procedure is the metropolis-hastings sampler discussed previously. Once a sample of graphs has been identified the estimation algorithm is launched. Since the normalizing constant in equation 2.27 cannot be compute a workaround has to be found. The idea is to give starting parameters (θ_0).The log-likelihood ratio can then be written (Hummel et al., 2012):

$$L(\theta) - L(\theta_0) = (\theta - \theta_0)^T V(G) - logE_{\theta_0}[exp(\theta - \theta_0)^T V(g)] \tag{2.28}$$

Geyer and Thompson (1992) point out that maximizing this ration by the means of a sample distribution of graphs generated with θ_0 only behaves well when θ is close to θ_0. In other words one has to choose the correct starting point for the algorithm to find the MLE.

The MLE solves the equation:

$$E_{\hat{\theta}} v(G) = v(G^{obs}) \tag{2.29}$$

The idea is to suppose that the MLE is not the observed value of the parameters but some point between the mean value parameterization and the observed value. A parameter γ defines the steps taken:

$$\hat{\omega}_t = \gamma_t \cdot v(G) + (1 - \gamma_t)\bar{\omega} \tag{2.30}$$

Where ω_t represents the estimate in the mean parameter space[5].Ideally then, we would want $\gamma = 1$ so that the expected value of the parameters is the observed value. If this is the case the algorithm is considered to converge, this is shown in figure G.1 which is the output of the R code. Once convergence is detected a large sample based on the parameters is computed and the MLE are estimated and gives as the final values.

Step 1 : Set the iteration number equal to 0 and choose initial parameters for vector η_0.

Step 2 : Use MCMC to simulate graphs from the probability function for parameter vector η_0.

Step 3 : Compute the mean of the sample.

Step 4 : Define a pseudo observation that is a convex combination of the mean of the sample and the observed value.

Step 5 : Replace the observed value by the pseudo observation.

Robbins-Monro The Robbins-Monro algorithme is a stochastic approximation method introduced by Robbins and Monro (1951) which is used by Snijders (2001) and Snijders (2002) to estimate ERGM models. Typically the method estimates:

$$E\{Z_\theta\} = 0 \tag{2.31}$$

[5]In the algorithm the initial values are chosen to be the MPLE

Where Z_θ is a vector of parameters equal to $u(Y) - u_0$ where u_0 is the observed value of the statistics. This allows us to rewrite the equation as a moment equation. The algorithm gives starting parameter values equal to the average of the parameter values. The initial parameters that will launch phase two are defined by(Lusher et al., 2012):

$$\theta^{t+1} = \theta^t - a_t \cdot D^{-1}(z(x^m) - z(x_{obs})) \tag{2.32}$$

Where D is the co-variance matrix, the diagonal of this matrix will be used as the scaling matrix. a defines the convergence, it is set to $a_t = \frac{a_{t-1}}{2}$. The idea is that each step brings the values closer to the MLE. Hence large steps might result in the exceeding the MLE and divergence. The fact that the a_r reduces in value with each step allows a smooth path to the MLE. As we move closer to the observed values of the statistics $z(x^m) - z(x_{obs})$ tends towards zero. The R output in figure G.1 shows how the steps (a) start at a value of 0.1 and tend towards 0 with each iteration of the second phase of the algorithm. At the start of each step the starting parameters are considered to be the average values of the previous step. The number of iterations varies from model to model. The iterations stop once the trajectories of the generated statistics cross those of the observed ones (Lusher et al., 2012).

The burn-in represents the number of simulations that are removed from the MCMC in order to make the chain "forget" the starting point. In other words it is to make sure the starting values do not impact the final result.

Finally the algorithm checks for convergence using a convergence statistic. Just as in the case of the stepping algorithm one supposes that the MLE is reached when the distance between the observed values and the average of the simulated ones is close to 0. If there is no convergence than one can relaunch the estimation with as starting parameters the results of the previous simulation (Lusher et al., 2012).

The largest difference between this method and the stepping method resides in two factors. First this method approaches an estimate of the MLE and does not evaluate the MLE function directly. Second, the steps are of a higher magnitude and can exceed the MLE if the starting values are close to the MLE. The use of either of the algorithms purely depends upon the model to be estimated. One algorithm might have better convergence in one case while the opposite can be true in another case. Note however that both use the

Metropolis-Hastings method for the simulation of the MC.

Now that we have discussed which variables can be included and how to estimate the parameters we will turn to an application using R.

2.4 Code R and example

We use here different R packages (Hunter, Handcock, Butts, Goodreau, Morris and Martina, 2008) , (Hunter, Handcock, Butts, Goodreau and Morris, 2008; Handcock et al., 2008; Butts, 2008).

The data (and hence the results of the estimations) are from the French aerospace collaboration network. Using patent data collaborations were identified which resulted in a network. The aim of the study is to understand if technological proximity played a significant role in the structuring of the collaboration network. We hence used a dyadic covariate called "proximity". The network contains 176 firms.

```
#Import data.
Network<-read.table("ADJ_MATRIX.csv", sep=";")
```

The data used here was already in the form of an adjacency matrix and hence could be used directly. It is however also possible to use edgelists. Since the data needs to be transformed into a network object the network package will be needed. The latter is able to transform edgelists into adjacency matrices.

```
#Import the dyadic covariates
proximity<-read.table("Proximity_matrix.csv", sep=";")
proximity.e<-read.table("Proximity_matrix_exp.csv", sep=";", dec=",")
citation<-read.table("Citation_matrix.csv", sep=";")
```

Since I'm using two measures of proximity I have two matrices and a matrix that includes the number of citations between firms. These need to be imported in the same manner as the network itself.

```
#We now transform the imported data into network objects with the
    package 'network'
Network<-as.matrix(Network)
proximity<-as.matrix(proximity)
proximity.e<-as.matrix(proximity.e)
```

```
#Transform to network format
Network<- as.network(Network, directed = FALSE)
proximity<-as.network(proximity, directed=FALSE)
exp_proximity<-as.network(exp_proximity, directed=FALSE)
citation<-as.network(citation, directed=TRUE)
```

We now have different objects to work with: the network and dyadic covariates in the form of networks. Note that the model constructed here is for un undirected network as shown by the option $directed = FALSE$ in the $as.network$ function.
We now have to decide which variables to include in the model. Let's start with a simple model containing only edges. We invoke here the $ergm()$ function from the $ergm$ package:

```
model<-ergm(Network ~ edges)
```

This gives us the most basic form of an ERGM model, the estimation method defaults to Monte Carlo MLE (Geyer and Thompson, 1992).

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges

Iterations:  7 out of 20

Monte Carlo MLE Results:
          Estimate      Std. Error  MCMC      \%p-value
edges  -3.85280     0.05649       0            <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 21349  on 15400  degrees of freedom
 Residual Deviance:  3113  on 15399  degrees of freedom

AIC: 3115     BIC: 3122     (Smaller is better.)
```

The parameter for the variable $edges$ has an estimated value of -3.8528. This means

that the probability of two ties connecting is:

$$p(i \rightarrow j) = \frac{exp(-3.85)}{1 - exp(-3.85)} = 0.02174241 \tag{2.33}$$

Recall equation 2.11, this equation stated that the variables were change statistics. The parameter should hence be multiplied by the change in the number of subgraphs. In other words, if an additional edge creates three triangles the parameter should be multiplied by three. Since an additional edge only creates one new edge we do not multiply. Let's try the same but with only triangles as explanatory variable.

```
1  ==========================
2  Summary of model fit
3  ==========================
4
5  Formula:   Network ~ triangles
6
7  Iterations:  NA
8
9   Stepping MLE Results:
10           Estimate Std. Error MCMC \% p-value
11 triangle -1.91496    0.01712       0  <1e-04 ***
12 ---
13 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
14
15      Null Deviance: 21349  on 15400  degrees of freedom
16  Residual Deviance:  8935  on 15399  degrees of freedom
17
18 AIC: 8937    BIC: 8945    (Smaller is better.)
```

The parameter for the variable triangles has an estimated value of -1.9146. This means that the log-odds of two nodes connecting is:

$$-1.9146 * \delta\ triangles \tag{2.34}$$

Where $\delta\ triangles$ gives the change in the number of triangles. Hence the log-odds depend upon the number of triangles that will be created by an additional tie:

> If the link creates 1 triangle, the log-odds are 1 * -1.9146. The probability is then 0.1284

> If the link creates 2 triangles, the log-odds are 2 * -1.9146. The probability is then 0.0213

Let's see how we interpret estimates when we have more than one variable: the Markov model.

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges + kstar(2) + triangles

Iterations:  NA

Monte Carlo MLE Results:
          Estimate Std. Error MCMC % p-value
edges     -28.30222    0.22246      0  <1e-04 ***
kstar2      7.68159    0.07356      0  <1e-04 ***
triangle   57.09972    0.79544      0  <1e-04 ***
---
Signif. codes:  0 '**' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance:    21349  on 15400  degrees of freedom
 Residual Deviance: 35811780  on 15397  degrees of freedom

AIC: 35811786     BIC: 35811809     (Smaller is better.)
```

In this case an additional edge, if it creates x 2-stars and y triangles, has log-odds:

$$-28.30 + x \cdot 7.68 + y \cdot 57.099$$

Since the model includes lower and higher order subgraphs (2-stars are a substructure of triangles) we can conclude here that triadic closure is significant in the network. In other words, 2-stars are closed to form triangles.

Note however the values of the information criteria AIC and BIC, stating these values are high is an understatement. The model must suffer from some kind of problem. Subsection 2.4.3 will show how to deal with degeneracy and other problems.

Now we want to see how proximity (technological in the case of this study) influences the structuring of the network. Let's start with a simple model using the edge covariate "proximity". The comment for adding an edge covariate is simple *edgecov*. Similarly node covariates (age, gender, type of firm, country...) are added using the *nodecov*() or *nodefactor* command. The node factor command is particularly useful since it allows to compare log-odds to a reference point. For example, one could categorize the R&D expenses of firms into *low*, *average* and *high*. The ERGM would then give the log-odd of 2 categories as compared to a third. In other words is a link more likely for firm with average expenses than for low ? This feature seems not to be available for edge covariates however.

The following model only contains edges and the edge covariate. A first point we can notice is that the AIC and BIC criteria are lower with the addition of the edgecovariate. The model is hence improved with the addition of the proximity parameter. Firms with proximity are 2.13 times more likely to connect in this network (citeris paribus)[6]. The probability of an additional edge is then positively impacted by the technological proximity. More specifically the average degree of the network is positively impacted by technological proximity.

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges + edgecov(proximity)

Iterations:  7 out of 20

Monte Carlo MLE Results:
                  Estimate Std. Error MCMC %  p-value
edges              -4.5222     0.1972      0  < 1e-04 ***
edgecov.proximity   0.7575     0.2058      0 0.000234 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 21349  on 15400  degrees of freedom
```

[6] $e^{0.7575}$ since this is the odds and not the log-odds

```
 Residual Deviance:  3096  on 15398  degrees of freedom

AIC: 3100     BIC: 3115     (Smaller is better.)
```

A network analysis performed on this network showed that the network has a scale-free structure. This information can be helpful in the modeling of the ERGM as we have information on the distribution of the degrees. The same is valid for any other information about the network, small-world properties, level of clustering or centrality distribution. The information provided by SNA allows a first understanding of the structural properties of the network that will allow for a more robust model once edge and nodal covariates are added.

So if the network ha a scale-free structure the structure should be explained by the degree distribution. To check this we can add different degrees to the model. This can be done by using the command $degree()$. On can simple add one statistic for one degree, i.e $degree(3)$, for the impact of nodes with degree 3, or add multiple degrees as was done in the following model. Note that the addition of multiple degrees is achieved by writing $degree(a:b)$ to add all the degrees between a and b.

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges + degree(2:6) + edgecov(proximity)

Iterations:  7 out of 20

Monte Carlo MLE Results:
                  Estimate Std. Error MCMC % p-value
edges              -4.2417     0.1606      0 < 1e-04 ***
degree2            -0.8945     0.2233      0 < 1e-04 ***
degree3            -1.4608     0.2381      0 < 1e-04 ***
degree4            -1.8040     0.2720      0 < 1e-04 ***
degree5            -1.6668     0.2699      0 < 1e-04 ***
degree6            -2.0959     0.4109      0 < 1e-04 ***
edgecov.proximity   0.5451     0.1726      0 0.00159 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
20
21      Null Deviance: 21349  on 15400  degrees of freedom
22  Residual Deviance:  2989  on 15393  degrees of freedom
23
24 AIC: 3003     BIC: 3056     (Smaller is better.)
```

The addition of these variables to the model once again decreases the AIC and the BIC, the model is hence enhanced. The structure of the network can be explained by a degree distribution.

```
1 ==========================
2 Summary of model fit
3 ==========================
4
5 Formula:   Network ~ edges + degree(2:6) + edgecov(proximity2)
6
7 Iterations:  7 out of 20
8
9 Monte Carlo MLE Results:
10                   Estimate Std. Error MCMC \% p-value
11 edges              -4.2382     0.1815      0 < 1e-04 ***
12 degree2            -0.8764     0.2053      0 < 1e-04 ***
13 degree3            -1.4438     0.2229      0 < 1e-04 ***
14 degree4            -1.7876     0.2728      0 < 1e-04 ***
15 degree5            -1.6763     0.2874      0 < 1e-04 ***
16 degree6            -2.1068     0.4595      0 < 1e-04 ***
17 edgecov.proximity2  0.5446     0.1941      0 0.00502 **
18 ---
19 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
20
21      Null Deviance: 21349  on 15400  degrees of freedom
22  Residual Deviance:  2989  on 15393  degrees of freedom
23
24 AIC: 3003     BIC: 3057     (Smaller is better.)
```

These models seem to work qui nicely. We discussed in the previous sections that models were prone to degeneracy and the solutions to this problem. Let's have a look at how we can model ERGMs with alternating k-stars and a geometrically weighted degree distribution.

2.4.1 Curved Exponential Random Graph Models

Studies show that the addition of weights on the degree distribution helps to avoid bi-modal distributions in the parameter space, i.e avoids the generated networks from being either full or close to empty. Different forms can be added to the R code. Since we have here an undirected network we can use either the alternating k-stars $altkstar()$ or the geometrically weighted degree distribution $gwdegree$. For directed graph there are additional commands which work in similar manner as what we show here. We include here a statistic for the gwdegree with the option $fixed = TRUE$. The latter means that we do not make an estimation of the scaling parameter, we want it to be equal to 1. The resulting model is hence not a curved ERGM.

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges + triangles + edgecov(proximity2) + gwdegree
    (1,
     fixed = TRUE)

Iterations:  NA

 Stepping MLE Results:
                      Estimate  Std. Error MCMC %  p-value
edges                -5.717e+00   1.829e-01      0  < 1e-04 ***
triangle              1.802e+00   3.026e-05      0  < 1e-04 ***
edgecov.proximity2    6.811e-01   2.159e-01      0 0.001607 **
gwdegree              2.917e-01   8.380e-02      0 0.000502 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 21349  on 15400  degrees of freedom
 Residual Deviance:  4170  on 15396  degrees of freedom

AIC: 4178     BIC: 4208     (Smaller is better.)
```

In order to have a curved exponential random graph model, the parameter that defines that we fixes in the previous code has to be estimates as well. In the following code we

estimated the model with an edgewise shared partners variable (see figure 2.24). This variable is used to check for transitivity. By switching the option $fixed = TRUE$ to $fixed = FALSE$ the model becomes curved. The results now include an estimate for the parameter alpha of the model. Note here that the parameter can only be interpreted if the *gwesp* statistic is significant.

```
==========================
Summary of model fit
==========================

Formula:   Network ~ edges + edgecov(proximity) + gwesp(alpha = 1,
    fixed = FALSE)

Iterations:  NA

 Stepping MLE Results:
                  Estimate  Std. Error MCMC % p-value
edges             -5.39343     0.45601      0  <1e-04 ***
edgecov.proximity  0.48255     0.51641      0    0.05 ***
gwesp              1.19503     0.08333      0  <1e-04 ***
gwesp.alpha        0.88784     0.09792      0  <1e-04 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 21349  on 15400  degrees of freedom
 Residual Deviance:  2847  on 15396  degrees of freedom

AIC: 2855     BIC: 2885     (Smaller is better.)
```

The interpretation of these estimates are much more complex than previously. The parameters need to be exponentiated to find λ that we saw in the equations (Hunter, 2007). We hence find $e^{0.88784} = 2.42$. Since the parameter is positive we can conclude that transitivity is present.

2.4.2 Goodness of fit diagnostics

In order to check if a model is a good fit we use the *mcmc.diagnostics* command. This gives us a number of outputs, notable the matrix of correlations and p-values for both

the individual parameters and the model as a whole.

```
Sample statistics cross-correlations:
                    kstar2 edgecov.proximity
kstar2              1.0000000          0.5494967
edgecov.proximity 0.5494967          1.0000000

Individual P-values (lower = worse):
           kstar2 edgecov.proximity
        0.3126642          0.9294963
Joint P-value (lower = worse):  0.6994233 .
```

The p-values are high for the parameters and the model, we can hence conclude that the model is globally significant. In order to go into a bit more detail when it comes to the estimates, the commands also provides us with plots, see figures 2.8 and 2.9. We stated that an ERGM should fit the observed network perfectly, on average. This means that from the simulated networks we expect the average values to be those of the observed network. If this is not the case then the sample the model is based on does not come from the same distribution as the observed network.

Figure 2.8 shows an example of what we want to observe. in the first graphs the values oscillate around the mean which is what we want. The graphs on the right hand show a centered distribution of the values, we hence conclude that the model is a good fit. A bad example can be found in figure 2.9. The graphs show that the distribution is not at all centered, and there is no oscillation around the mean. This model is hence a bad fit.

One can also study the goodness of fit using the $gof()$ command. Using a plot command this provides a box plot (see figure2.10).

The gof command can receive different parameters, one can chose to plot any number of variables and decide to increase or decrease the number of simulations to refine the results. The following code provides the GOF of the whole model ($GOF = Model$) using 20 simulations ($nsim = 10$).

```
gof_model<-gof(model, GOF=~Model, nsim=20)
```

The results give us a box plot per variable and a black line representing the observations on the empirical network. Since we want the mean of the simulations to be equal to the observed network, the dark line should coincide with the center of the boxplots (the vertical line in the boxplot representing the median of the distribution). This is the case here, the

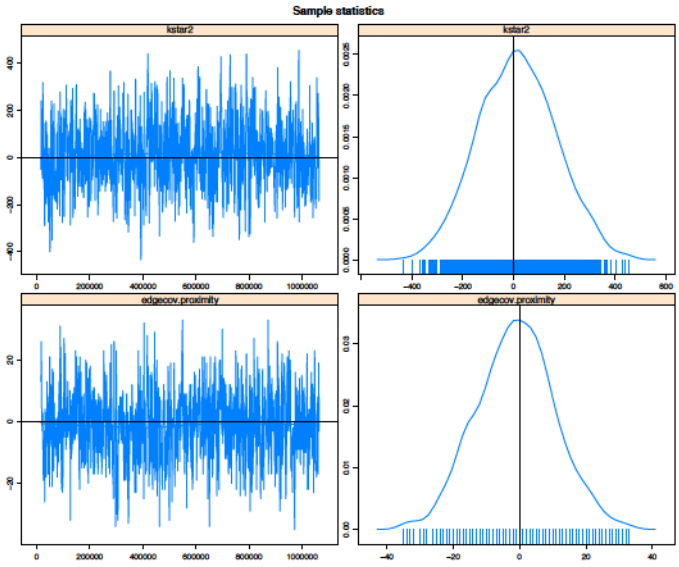


Figure 2.8: *Goodness of Fit diagnostics*

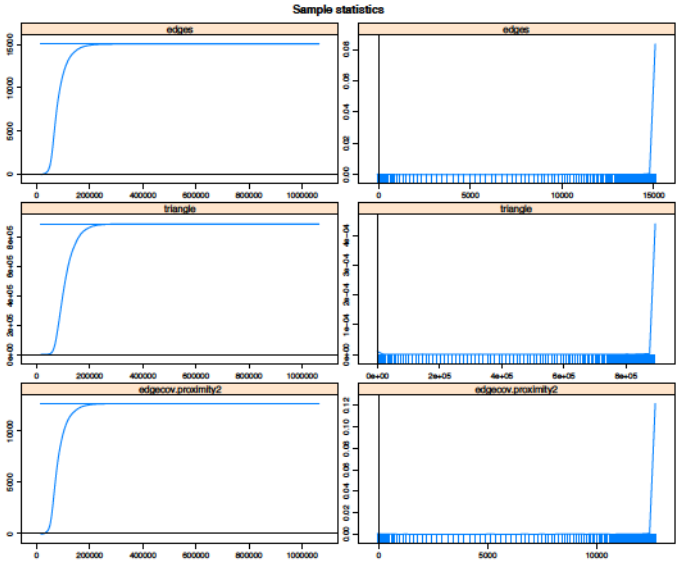


Figure 2.9: *Goodness of Fit diagnostics, bad example*

model is hence a good fit.

2.4.3 Improve bad models

The fitting of an ERGM is a trial and error procedure. If a model behaves badly there are a couple of parameters to change in order to improve results. Of course one should only start these procedures once the variables chosen are stabilized. Starting with a null model containing only edges and adding on to this model while comparing the AIC and BIC values to find the variables of importance.

Once this is done and degeneracy is still observed one can start by switching estimation methods. One method might work better in one case than the other.

The estimation algorithm can be chosen in the control arguments of the $ergm()$ command.

```
model<-ergm(Network~ edges, control=control.ergm(main.method = "
    Stepping"))
model<-ergm(Network~ edges, control=control.ergm(main.method = "
    Stepping",MCMC.samplesize=70000, MCMC.interval=5000))
```

The burn-in can also solve problems, the burin represents the number of iterations that are removed from the simulation. It other words, the higher the burnin the more the procedure forgets about it's initial parameters. Increasing this value hence allows for keeping only the latest values which might represent the real values better. This can be achieved by adding an option to the ergm.

A second method of improving estimation would be to increase the sample size by changing the parameter $MCMC.samplesize$. This increase will result in having more precise estimates by an increase in the number of statistics drawn from the sample. This, of course, increases computation time. The ergm package includes multicore features that can help reduce computation time drastically. All this requires is the addition of some parameters to the $control.ergm$ argument. Adding $parallel = 4$ notifies the package that the computer has 4 cores, $parallel.type$ sets the type of multicore. For a regular computer this should be fixed to "$PSOCK$". This will distribute the computations over the 4 cores of the computer and hence increase speed.

```
a<-ergm(Network~edgecov(proximity2)+triangles+altkstar(1.812, fixed=
    FALSE),control=control.ergm(main.method=c("Stepping"), parallel=4,
    parallel.type="PSOCK", MCMC.samplesize=20000))
```

2.5 Conclusion

ERGM models are able to analyze the structure of network to an extent that other methods of network analysis are unable to reach. A basic analysis of the network structure can come in handy when defining the ERGM model and can be used as a verification procedure. Whether they are used to analyse social networks, collaboration networks, trade networks or financial networks, ERGMs can provide vital insights into the understanding of network dynamics. Even though the models are powerful, the tools used for their

analysis still need to be improved. The tweaking required in the estimation procedures should be reduced for the models to be used by a larger audience.

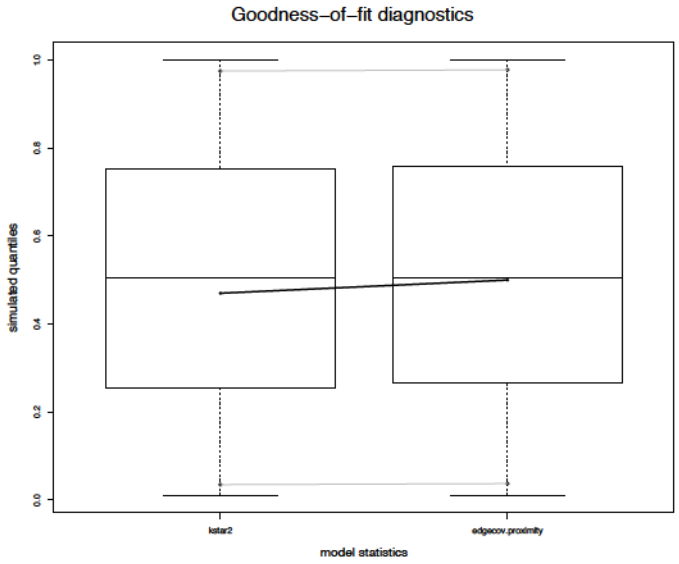


Figure 2.10: *GOF: boxplot analysis*

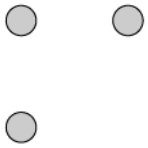


Figure 2.11: *Empty*

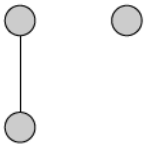


Figure 2.12: *Dyad*

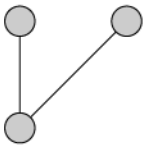


Figure 2.13: *2-star*

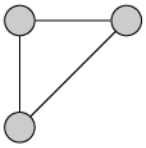


Figure 2.14: *Triad*

Figure 2.15: *Triads*

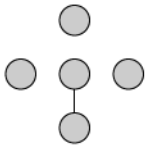


Figure 2.16: *1-star*

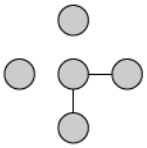


Figure 2.17: *2-star*

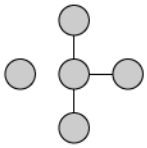


Figure 2.18: *3-star*

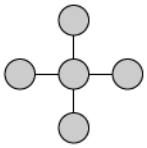


Figure 2.19: *4-star*

Figure 2.20: *k-stars*

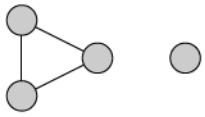


Figure 2.21: *1 partner*

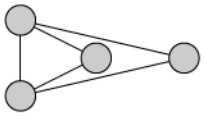


Figure 2.22: *2 partners*

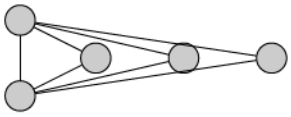


Figure 2.23: *3 partners*

Figure 2.24: *Shared partners (edgewise)*

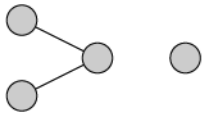


Figure 2.25: *1 partner*

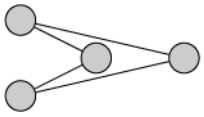


Figure 2.26: *2 partners*

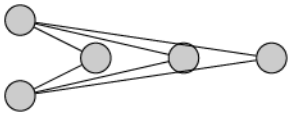


Figure 2.27: *3 partners*

Figure 2.28: *Shared partners (dyadic)*

Part II

Part B: Network dynamics and the impact of structure on performance

Chapter 3

The co-evolution of knowledge and collaboration networks: the role of the technology life-cycle in Structural Composite Materials

"This is one corner... of one country, in one continent, on one planet that's a corner of a galaxy that's a corner of a universe that is forever growing and shrinking and creating and destroying and never remaining the same for a single millisecond. And there is so much, so much to see" - The Eleventh Doctor

3.1 Introduction

In this chapter I focus on the impact of a specific factor on the structural dynamics of a collaboration network. The structure of a collaboration network is influenced by a multitude of factors. Factors such as industry (Salavisa et al., 2012), types of actors included (Nieto and Santamaría, 2007) as well as geography (McKelvey et al., 2003) have shown to have an important impact. However, the role played by the technology life-cycle is still mostly unexplored (Stolwijk et al., 2013). This chapter extends the existing literature on innovation networks by analyzing the impact of the life-cycle of the technology on the structural dynamics of the network.

Technologies are developed in different phases. Broadly speaking, there is a research phase and a development phase. During the first phase fundamental research is performed while

the second phase is aimed at developing applications for the technologies. It would then be reasonable to hypothesize that different firms enter the network during each of these stages. The first stage would reveal the inclusion of research institutions, while the second phase would require more market oriented knowledge coming from firms. The question I will try to answer in this chapter is how this life-cycle defines the structure of the collaboration network around this technology. The effect of the life cycle is really only visible when analyzing the collaboration network around a specific technology. At higher levels of aggregation such as sectors and regions, a multitude of technologies co-exist and it would be difficult or even irrelevant to analyze the impact of a life-cycle at those levels. For this reason I focus here on a specific technology that relates to a concern as old as aviation itself: weight reduction. Makes aircrafts lighter reduces fuel consumption and hence reduced the environmental impact of the aircraft. To reduce the weight of airplanes, firms have focused on the implementation of composite materials in the structural elements of aircrafts. The technology chosen for this chapter is therefore Structural Composite Materials (SCM) in aeronautics. The case is particularly interesting since composite materials have been around for a long time. The first composite material, plywood, was created in Mesopotamia in around 3200 B.C. Even though the concept of composte materials is still the same, the technologies required for the production of the modern versions of composite materials are highly complex and new applications require new research and development.

Before being applied to aeronautics, modern composite materials were used in the automotive industry. Firms in the aerospace industry could use what has been done in the automotive industry but faced specific challenges that required further research and development. For instance, the conductivity of the structure in the case of a lightning strike becomes a major issue since CM are much less conductive than the steel counterpart. In addition there were production issues. The size of the components from an airplane are not comparable to those of parts in a formula one race car. The production of large sheets of composite materials using the same techniques was impossible. Furthermore, the larger the sheets, the less rigid the product. Airplanes however, require rigid structures.

The analysis of the technology then shows how firms in the aerospace sector researched and developed this technology so that it could be applied to structural components of aircrafts. This allows me, in addition to the first research question, to compare two diverging strategies in terms of knowledge absorption. Airbus and Boeing both aimed at implementing

the technology into their airplanes but applies different strategies. One researched and developed the technology with its historic partners while the other sought the help of firms that had developed the technology in other sectors.

3.2 Literature and hypotheses

3.2.1 The technology life-cycle

We aim at identifying a correlation between the evolution of the collaboration network and the technology life-cycle. Patent and publication data are widely used for the analysis of the technology life-cycle (Alencar et al., 2007; Gao et al., 2013; Trappey et al., 2013), however not in network form. We start from the Schumpeterian perspective that innovations are the result of the recombination of existing technologies. International Patent Classification codes (IPC codes) present on patents can be used as a proxy for "technology blocks". The presence of different codes on the same patent bares witness to a recombination of technologies. By taking all the IPC codes present on patents a network is created. We use the network created by the IPC codes as a proxy for the identification of the evolution of the life-cycle of the technology. Our conjecture is that during the first stage of the life-cycle, fundamental knowledge about the technology is researched. The deposited patents will be characterized by a relatively small number of IPC codes. The patents deposited during this phase form the core of the technology. The low number of codes and high number of deposits result in an IPC network that will be very densely interconnected. From a dynamic perspective we should observe a continuous increase in the clustering coefficient of the IPC network over time.

Hypothesis 1: *The clustering coefficient of the IPC network will increase continuously during the research stage.*

As the technology moves from the research phase to the development phase, patents deposited for incremental innovations combine the core technology with new fields of application. This results in the inclusion of new technology blocks in the network.

The developments on the core technology are relative to a specific application of the technology. We can take the example of a photo camera. The core technology is the

camera, an application would be the integration of the camera into laptops, phones and watches. Each application has a specific research direction. As a result, the new blocks connect to the core but only scarcely connect between them. From a structural point of view new nodes are added to the network, creating a periphery around the core created by the research stage. This bring us to our second hypothesis:

Hypothesis 2: *During the development phase, the clustering coefficient decreases continuously due to the addition of nodes in the periphery. This has the associated result of increasing the average distance in the network.*

As the development stage evolves the periphery develops while at the same time reinforcing the core of the technology. This leads us to our third hypothesis:

Hypothesis 3: *The knowledge network has the structure of a core-periphery network.*

3.2.2 The collaboration network

Understanding the structure of a network is highly dependent upon the environment in which it evolves. The collaboration network of the aerospace sector does not have the same structure as the collaboration network of the biotech sector. The first is based on a highly optimized production chain, while the second is a highly competitive horizontal sector. The focus of this chapter is on a collaboration network at the level of one specific technology (SCM) inside an existing sector. At this level the life-cycle of the technology intervenes as a defining factor in the structure while it does not at the level of the sector. Since many life-cycles evolve continuously at different stages this would be difficult to track.

We follow the definition of a technology life-cycle based on two stages, a research stage and a development stage (Davide Chiaroni, 2008; Rowley et al., 2000; Virapin and Flamand, 2013). The first stage is characterized by a research phase in which fundamental knowledge is required to create a new technology. This stage requires collaborations with agents that possess fundament knowledge and are able to conduct basic research. Once the technology has been stabilized, the development stage begins. During this stage, firms start to apply and develop their technology for different applications. During this phase new

collaborations are required with other agents with new abilities in other fields. Different applications for the technologies will be developed by different clusters of firms each with their own specialization. A dense interconnection of firms for the basic research of the technology is then expected to appear in the early stages of the network. For this reason we expect the network to exhibit high levels of clustering.
As times goes by, clusters of firms developing applications for the technology will connect to the initial cluster of collaborations during the second phase. This interconnection of clusters should result in a low average distance between firms. Taken together, the high clustering and low average distance represent a global network structure that is referred to as a "small world". This gives us our fourth hypothesis.

Hypothesis 4: *The structure of the collaboration network converges towards a small world structure.*

Since the types of collaborations change around the same time as the stages of the life-cycle we formulate the following, final hypothesis:

Hypothesis 5: *The structure of the collaboration network is correlated with the life-cycle of the technology.*

3.3 Data and methodology

3.3.1 Structural Composite Materials

Structural Composite Materials (SCM) were first developed by chemists in the early 20th century and have since been used in sport equipment and the automotive industry (Virapin and Flamand, 2013). It caught the attention of civil aircraft manufacturers during the late 70's. During this period, research programs focusing on the optimization of energy consumption were launched by the European Union and the american government. The aim of these programs was to exploit composite materials in order to increase energy efficiency for aircrafts by the means of weight reduction. This makes SCM the perfect candidate for a study to analyze how a network is structured in order to absorb an existing technology from other sectors and develop it for its own needs.

The aerospace sector has a particular structure, it is organized as a production chain. An aircraft being a multi-technological product, each part of the airplane is developed in a different part of the network.

In the value chain that makes up the sector, a small number of firms occupies a strategic (central) position, these firms assemble intermediary products before sending them to either the final assembler (Airbus, Boeing etc.) or to other firms that use intermediary goods for larger parts. Firms with these specific positions in the value chain are called "pivot firms" (Frigant et al., 2006). These firms have to master all the technologies of the downstream firms in order to complete their part of the aircraft (Van Der Pol et al., 2014).

The introduction of a new technology such as SCM, can only succeed if the value chain adapts to the technology. Indeed, the introduction of SCM alter the structure of an aircraft in many dimensions. Pivot firms have to adjust their production methods and hence so do the downstream firms. Integrating SCM in the aerospace industry hence implies a thorough understanding of the core and linkage technologies (Prencipe, 1997) by all the actors implicated in aeronautical programs.

We used patent and publication data to generate our networks. Patents were extracted from Orbit while publications were extracted from Web Of Science (there were no geographical restrictions).

In order to extract all relevant patents and publication we started by framing the technologies involved in the production of SCM. The framing process is an iterative process based on discussions with engineers and executives from the aerospace sector.

We conducted an initial search for relevant IPC codes by identifying parts of the aircraft that can be made out of SCM. A detailed search was then conducted (combing IPC codes and key-words) in order to identify which specific products and technologies are involved in the creation of composite materials (resins, matrices). We then discussed these codes and key-words with engineers who would confirm or infirm the relevance. New codes and key-words were identified based upon these discussions and then discussed again. This iterative process allows us to frame the technology and build up a query that extracts patents beyond the scope of keywords which would result in false positives and unidentified patents and publications. The final query includes both relevant IPC codes and keywords.

A total number of 15313 patents and 9030 publications were identified worldwide between 1980 and 2014. The analysis was initiated in 1980 since it is the point at which

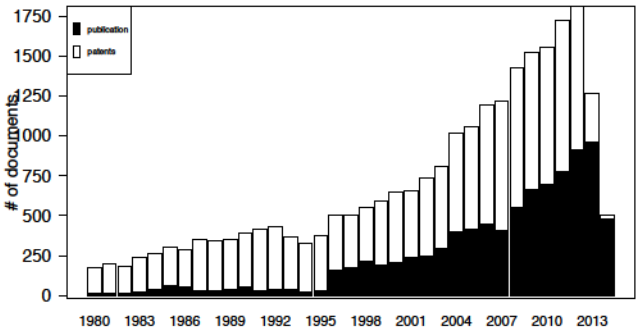


Figure 3.1: *Distribution of the number of patents and publications between 1980 and 2014*

SCM caught the attention of aircraft manufacturers. We checked patents and publication before 1980 and confirmed the latter. The results in figure 3.1 show the evolution of the number of patents and publications identified.

The Orbit database uses algorithms to extract and translate data from patents, this results in terms that get lost in translation and textual mistakes. In addition to this, names on patents often do not match names on publication. For example, we observe the name "Airbus SA" on a patent, "Airbus S.A" on a publication, even mistakes like "Aerhjbus" appear in the data. The entire dataset was cleaned by hand in order to ensure maximum accuracy of the results.

3.3.2 Methodology

Core-periphery (CP) identification

A CP network has a small number of highly connected nodes (the core) and a large number of (relatively) less connected nodes (the periphery). In other words, it is an interconnection of hubs. By representing the network by a Cumulative Frequency Distribution (CFD) of the number of links we can visualise a network and check for a Core-Periphery structure. A CFD is simply a plot with the frequency of nodes with degree k on the y-axis and the degree on the x-axis. This distribution is then transformed into a cumulative degree distribution as can be seen in figure 3.2. Figure 3.2 represents the CFD of the IPC network. One can see that roughly 95% of all nodes have at least two links, 80% has at least 3 links, and so on. If the frequency decreases by a small factor between densities, only a few nodes

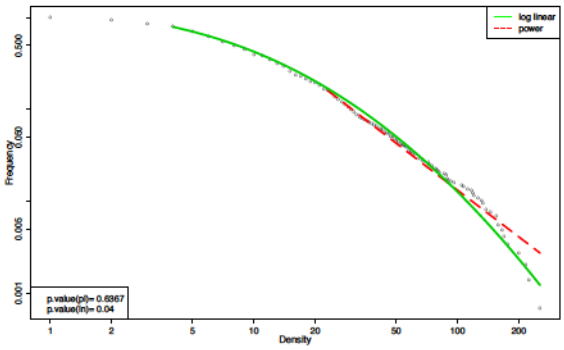


Figure 3.2: *Exemple of a Cumulative Frequency Distribution of the IPC network at the 7-digit level. The circles represent the frequency for each value of the density. The lines represent function that are fitted to the data.*

are lost due to the increase in density. If this CFD has the form of a line then when the degree increases by one, the frequency decreases by a fixed factor. The network is then called a scale-free network (since the diminishing factor is constant). This structure is represented by a power law which has the form: $p(k) = c \cdot k^{-\alpha}$. We also check for another form which is the log-normal function ($ln(k) = \frac{1}{k}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{ln(k)-\mu}{\sigma})}$). The main difference between the two functions is that they do not represent the same type of network. The scale-free network is a particular form of a core-periphery structure in which the frequency decrease is constant. In other words, when the density increases by one, the frequency drops by a factor k for all values of the density (it has hence the form of a straight line). Other functions such as the log-linear function can have more of a curvature to them. If the function is concave (figure 3.2) the drop in frequency increases with each increase in the density. The periphery of such a network contains less nodes with a low density. The periphery is less interconnected than the scale-free network. The inverse would be true if we were to have a convexe function. The shape of the adjusted function informs us about the type of core-periphery structure, ranging from sparse to dense.

In order to conclude to a core-periphery structure we fit a particular function to the data. The functions are fitted using a maximum likelihood estimation. We then use a bootstrapping method in order to assess the goodness of fit which provides us with a p.value. The null hypothesis (data comes from a power-law) is rejected when the p.value is below a fixed value.

Small World identification

Since we expect to find a small-world structure for the collaboration network, a method for the identification of such a structure is required. Small world structures are empirically important because they have features that favor the flow of knowledge between the firms inside the network. One of the reasons networks are important for innovation is the precisely the hypothesis that knowledge flows between firms. The structure of the network has an important role to play in the transfer of this knowledge. For knowledge to flow quickly through a network firms need to be a low distance from each other in the network. The higher the average distance in the network, the longer is takes for knowledge to reach all nodes. The latter is a necessary condition, however it is not sufficient. The presence of small communities is also a condition. Within these communities knowledge flows even faster since firms are closely connected to a larger number of other firms. New knowledge developed in these communities spreads fast throughout the community and then to the whole network. A network structure that has both the characteristics of low averagedistance and high clustering is a structure called the small world structure. It has been found to be an efficient structure for the diffusion of knowledge through a network(Cowan and Jonard, 2007; Verspagen and Duysters, 2004). We hence want to know if our network has the particular structure of a small world. In order to check for small world features we need information on the average distance in the network as well as the clustering coefficient.

The clustering coefficient is a measure of cohesiveness in a network, in other words, how well connected the network is. The measure is quite simple; it represents the number of triangles in the network divided by the number of possible triangles.

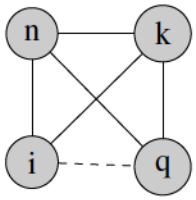


Figure 3.3: *Clustering illustration*

Consider figure B.2, to find the clustering coefficient we need the number of triangles in the network. There are two triangles in the network: i-n-k and n-q-k. The number of possible triangles is equal to the number of triangles if the network were a complete network. The dotted link between nodes i and q makes the network a complete

network. If this link existed we would have two additional triangles: i-n-q and i-k-q. The number of possible triangles is hence equal to four. The clustering coefficient is then equal to:

$$Clustering = \frac{\sum_{i,j\neq i,k\neq j,k\neq i} g_{ij} \cdot g_{ik} \cdot g_{jk}}{\sum_{i,j\neq i,k\neq j,k\neq i} g_{ij} \cdot g_{ik}} = \frac{2}{4} = 0.5 \tag{3.1}$$

The same value can be computed at the node level. This would give a measure of the extend to which firms' neighbors are connected. It gives the fraction of the neighbors that are connected.

Whether measured at the level of the node or the network level, the clustering coefficient gives a measure of embeddedness. When clustering equals one all possibles triangles exist, the more it tends towards zero the less triangles are observed.

When a network is studied in a dynamic setting a problem arises with this measure. The addition of nodes to the network result in an increase of the number of possible triangles, resulting in a reduction in clustering (we suppose that new firms do not collaborate with every other firm in the network but only with a small portion). If we simply compute the clustering coefficient it would be ever declining, it would hence not be a very useful measure. A method is required that allows us to measure if the network gets more clustered even when new nodes are added. This can be achieved by using a benchmark to compare the observed clustering coefficient(Watts and Strogatz, 1998; Baum et al., 2003; Gulati et al., 2012) to. The observed network is compared to a random network with the same number of links and nodes as the empirical network. Random networks typically have very low clustering since there is no reason why triangles would form a random. Clustering is then defined by the ratio :$\frac{C}{C_r}$. We use the same method for the average distance $\frac{L}{L_r}$.

Given that random graphs have low to no clustering and a low average distance we want a small world to show:

$$\frac{C}{Cr} >> 1 \qquad and \qquad \frac{L}{Lr} \approx 1 \tag{3.2}$$

3.3.3 Network dynamics

In order to track the evolution of the networks we use two different methods. For the collaboration network we start with all patents and publications that appeared in 1980

and extract collaborations. The latter is done by creating a link between firms that have co-deposited a patent or co-published a scientific article. Then, for the next year we add the collaborations that appeared that year. The network in 1985 hence contains all collaborations between 1980 and 1985. Co-patenting and co-publication collaborations are treated equally in the network.

In order to identify the life-cycle of the technology we use IPC codes present on the patents. Whenever two or more IPC codes are present on the same patent a link is created between them. The International Patent Classification (IPC) classifies patented technologies according to different technological fields. The system itself is crescendo is nature. The more digits, the more precise the codes become from a technological point of view. A 4-digit level defines a broad definition of technologies, for instance B64C defines "Aeroplanes; helicopters". The 7-digit level goes one step further into detail by specifying for B64C1, " Fuselages; Constructional features common to fuselages, wings, stabilizing surfaces, or the like". Going a step further, at the 9-digit level we will find "Floors" for code B65C1/18.

We use 3 different digit levels for our analysis, the 4, 7 and 9 digit levels. In order to obtain a 4-digit network, all IPC codes are reduced to their 4-digit format. For example B64C001/23 is a 9-digit code, in order to obtain the corresponding 4-digit code, one simply reduced the code to 4-digits: B64C. The same goes for the 7-digit code which would be B64C001 in this example.

Because of the hierarchy in the classifications, the 4-digit network will represent the interconnection of broad technological domains, while the 9-digit network pertains to more precise technological applications. The best fit for our analysis is hence the 9-digit network. However, the 4-digit network should have some interesting characteristics, it should show how different technological domains are interconnected.

3.4 Results

3.4.1 Structure of the knowledge network

The technology life-cycle

The 4-digit knowledge network represents the interconnection of broad technological fields. As shown in figure 3.4, the number of codes present in the knowledge network increases each year. However, each year less codes are added to the network. The number of links however, increases steadily over the same period of time. This shows that the major areas for the application of the technology are identified early on. New recombinations are however found between the technological fields. The high level of clustering shown in figure 3.10 shows that there was a dense interconnection of the different fields from the start. The fields that are added over time do not reinforce the core but rather expand it. The average distance between the nodes in the network hence increases as shown in figure 3.11. We notice here that the 4-digit network takes the particular structure of a small world. Figure 3.10 shows an adjusted clustering coefficient exceeding 1, even though declining over the period while figure 3.11 shows an adjusted average distance around the value of 1 around 1990. From that point on the network takes the structure of a small world and remains a small world during the rest of the period. This shows that the technological fields are locally clustered while all being at a close average distance from one another. Figure 3.16 shows the core of the network (the nodes with the highest number of links) The colors represent different communities in the complete 4-digit network. The network shows that even though these fields are densely interconnected they all can be divided into different communities.

More precisely, during the first decade (1980-1990) the IPC codes that had the largest number of deposits can be found in table 3.1

These codes relate to the fundamental development of the technology. In the succeeding decades, developments of the technology are added, baring witness to the start and evolution of the development phase. During these years, the IPC codes have changed to more diverse applications of the technology as shown in table 3.2.

The latter clearly show a switch from the research phase to the development phase of the technology.

IPC code	Definition
C08L	COMPOSITIONS OF MACROMOLECULAR COMPOUNDS.
C22C	ALLOYS
C08F	MACROMOLECULAR COMPOUNDS OBTAINED BY REACTIONS ONLY INVOLVING CARBON-TO-CARBON UNSATURATED BONDS
C07D	HETEROCYCLIC COMPOUNDS
C22F	CHANGING THE PHYSICAL STRUCTURE OF NON-FERROUS METALS OR NON-FERROUS ALLOYS

Table 3.1: *Table of the IPC codes with the highest frequency during the research phase*

IPC code	Definition
B82Y	SPECIFIC USES OR APPLICATIONS OF NANO-STRUCTURES
B82B	NANO-STRUCTURES FORMED BY MANIPULATION OF INDIVIDUAL ATOMS
D07B	ROPES OR CABLES IN GENERAL
B29C	SHAPING OR JOINING OF PLASTICS
B64C	AEROPLANES; HELICOPTERS
B64D	EQUIPMENT FOR FITTING IN OR TO AIRCRAFT
F02C	AIR INTAKES FOR JET-PROPULSION PLANTS; CONTROLLING FUEL SUPPLY IN AIR-BREATHING JET-PROPULSION PLANTS

Table 3.2: *Table of the IPC codes with the highest frequency during the development phase*

When we increase the number of digits, the precision of the technological fields increases. As a result, the core of the technology takes longer to stabilize. The 7-digit network stabilizes around the year 1990 while the 9-digit network stabilizes around the year 2000. For the identification of the stages of the life-cycle of the technology the 9-digit network is the most adequate. It represents the most detailed information about the domain of the technology and can hence be used for the identification of applications .

Figures 3.12 and 3.14 show an increase in the level of clustering from the beginning of the period. This represents the first phase of the technology life-cycle: the research phase. The fundamental technologies are interconnected until creating the core of the technology. Since an additional link increases the overall clustering of the graph we can deduce that the core is being reinforced as long as clustering increases. This observation is reinforced by figures 3.13 and 3.15 that show a decrease in the average distance of the network during that first phase. Technologies are interconnecting densifying the core, reducing the distance

separating them. From these observation we can confirm hypotheses 1 and 2.

The second stage of the technology life-cycle starts once the clustering coefficient of the network starts decreasing. New technologies are added to the network but they are not reinforcing the core, they stay in the periphery. The latter results in a decrease of the clustering coefficient and increases the average distance of the network. The codes that are added in the periphery of the network are applications of the technology. They do no connect to all the different IPC codes but rather to a specific fraction.

The increase of the average distance is difficult to observe in the 9-digit network which shows a stabilization of the average distance rather than an increase. In order to check the hypothesis that the structure is indeed a core-periphery structure we will use a statistical test on the degree distribution of the network.

Core-periphery identification

Figures 3.17 - 3.22 contains the degree distribution for the 7-digit and 9-digit networks. The green (plain) line is the fitted log-normal distribution, the red (dotted) line is the power law fit. The lower left corner of the graph contains the p.values. When the p.value is lower than 5% we reject the null hypothesis and conclude that the degree distribution has a core-periphery structure.

A first observation is that we reject the power-law fit for both the 7-digit and the 9-digit networks. The networks are not scale-free.

The log-normal fit is not significant in the first years of the 7-digit network, the fit becomes significant in 1995 (p.value = 0.27), and remains significant until the end of the period. Towards 2009, the parameters of the log-normal fit stabilize. Recall that the parameters of the log-normal distribution are the average and the standard-deviation. The variance tends towards a value of 1.96.

The 9-digit network appears to have a core-periphery structure quite early on, but the structure is not stable. The log-normal fit implies that the difference between the core and the periphery is less clean-cut as would be the case in a scale-free structure. The periphery is quite densely connected. The parameters of the network stabilize around the year 2000, the variance stabilizes around 3.24.

These observations allow us to conclude that the knowledge network for SCM technologies takes the structure of a core-periphery network and hence validate hypothesis 3.

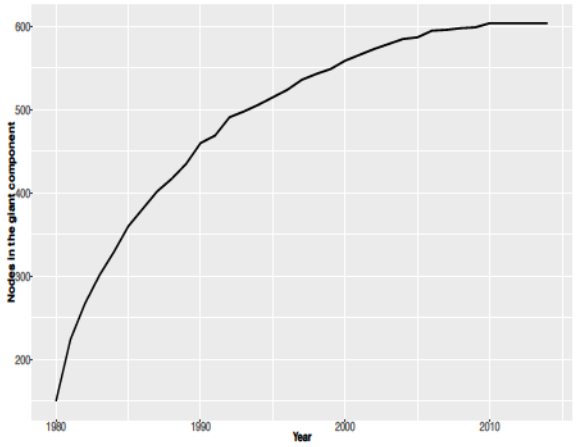


Figure 3.4: *Evolution of the number of nodes in the 4-digit IPC network*

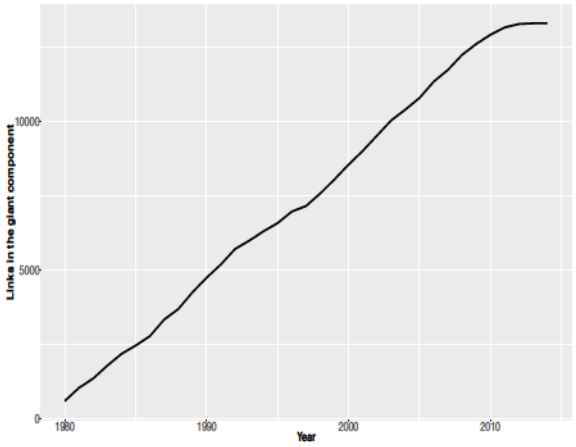


Figure 3.5: *Evolution of the number of links in the 4-digit IPC network*

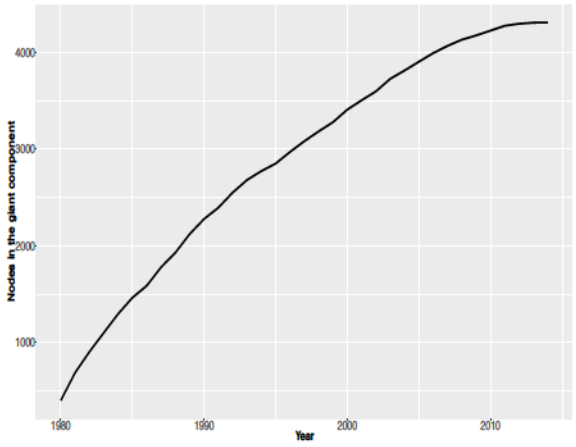


Figure 3.6: *Evolution of the number of nodes in the 7-digit IPC network*

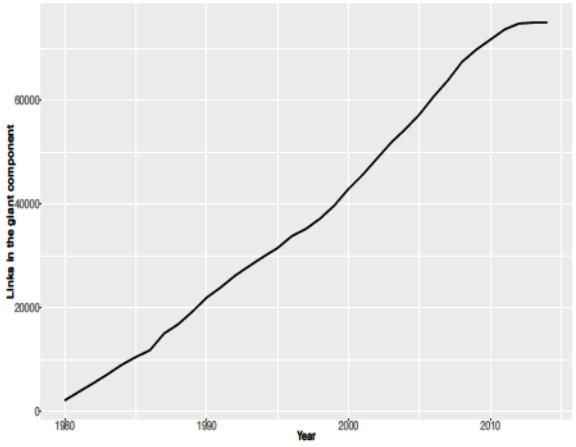


Figure 3.7: *Evolution of the number of links in the 7-digit IPC network*

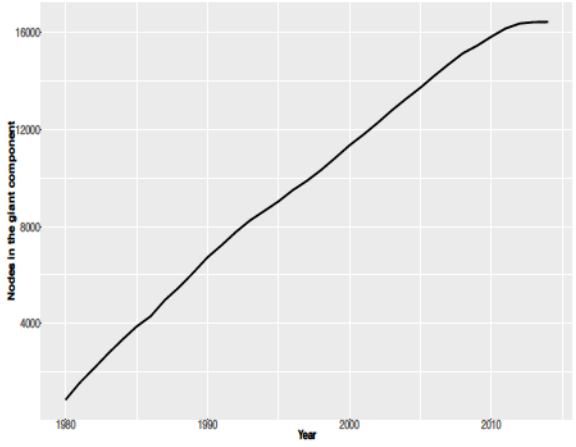


Figure 3.8: *Evolution of the number of nodes in the 9-digit IPC network*

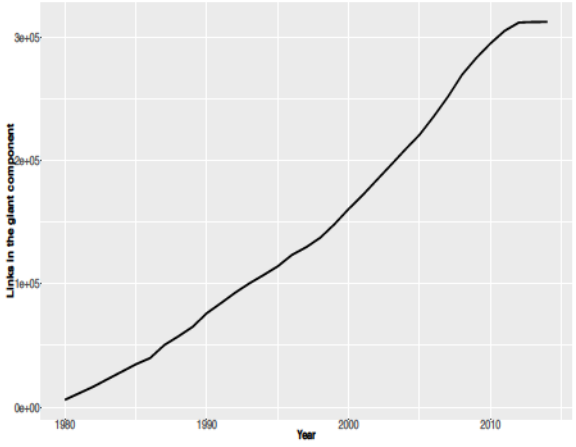


Figure 3.9: *Evolution of the number of links in the 9-digit IPC network*

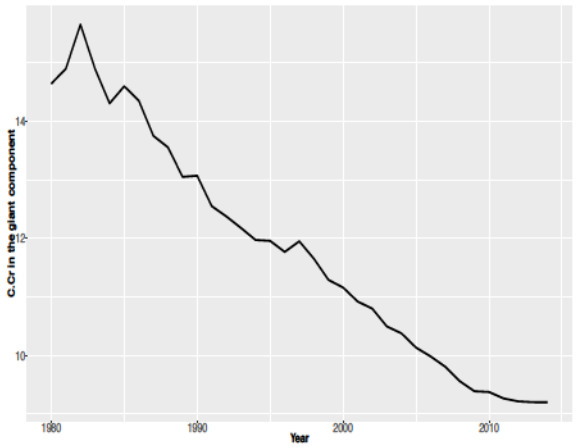


Figure 3.10: *Evolution of the adjusted clustering coefficient in the 4-digit IPC network*

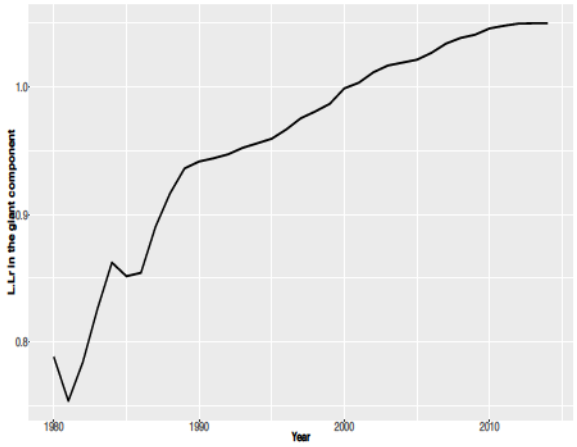


Figure 3.11: *Evolution of the number of adjusted average distance in the 4-digit IPC network*

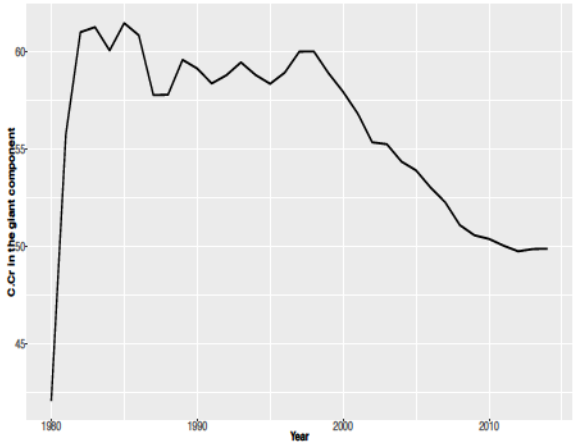


Figure 3.12: *Evolution of the number of the adjusted clustering coefficient in the 7-digit IPC network*

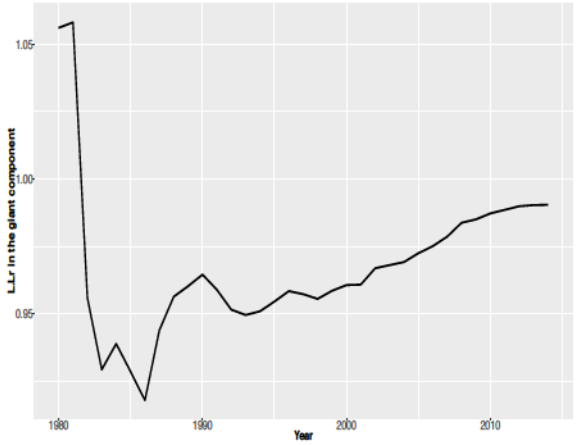


Figure 3.13: *Evolution of the adjusted average distance in the 7-digit IPC network*

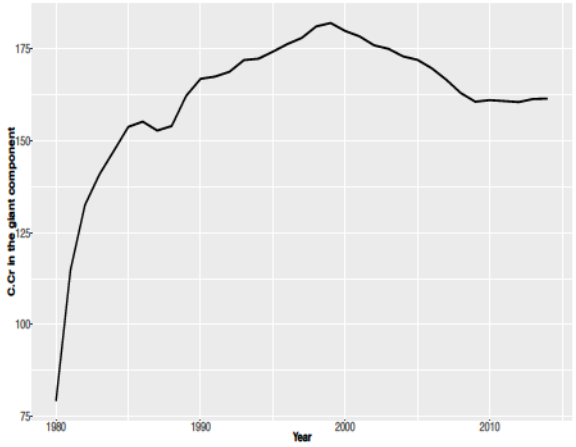


Figure 3.14: *Evolution of the adjusted clustering coefficient in the 9-digit IPC network*

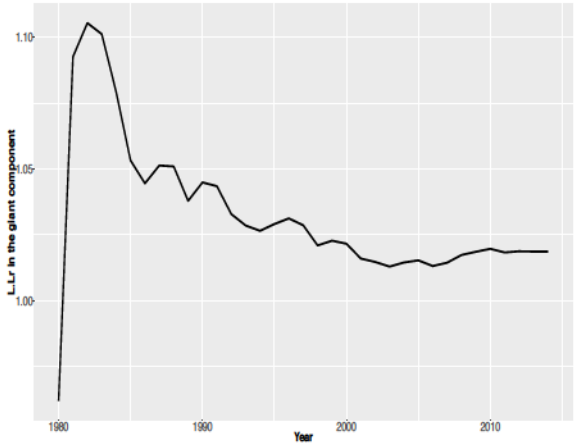


Figure 3.15: *Evolution of adjusted average distance in the 9-digit IPC network*

We have now identified the different stages of the technology life-cycle as well as the stabilization of the network structures.

3.4.2 Structure of the collaboration network

Since we use publication data and patent data we start by identifying the structure of each type of network separately before turning to the complete network. Publications have a higher average number of co-assignees than patents do. Performing the analysis on the separate networks allows a better understanding of the structure of the network as a whole.

The plots in figures 3.17-3.22 show that the patent network is structured early on in the analysis. The number of nodes and links is computed on the the largest component of the network. The publication network is build up from a large number of very small components that start to interconnect just before the year 2000.

Around the year 2008, the number of nodes in the publication network exceeds that of the patent network. The number of links is exceeded in 2005. In terms of links and nodes there is no clear cutoff point that indicates a switch from the research phase to the development phase. When we go into more detail we notice that there are many publications during the years 1980-1997 (see figure 3.1). The authors of these publications are actors of the space and defense sector that published mostly alone. The patents were deposited by the same type of actors (Boeing, US air force, Lockheed, Aerospace, NASA) but with collaboration. When we check the IPC codes on these patents we notice that they relate to the fundamentals of composite materials. In this case then, the research phase of the technology was accomplished by the private sector instead of the Universities.

The small number of collaborations in the publication network results in a low average distance as shown in figure 3.26. The average distance in the patent network is higher showing a larger diversity in firms. The distance increases over time because of new actors entering the network. In addition smaller clusters start to interconnect. Around the year 2000 the average distance in the patent network stabilizes. Less firms deposit patents while the number of publications increases. Actors from academia are developing the technology, the number of universities entering the network increases at a steady pace over that period of time.

The stabilization and decline of the patent network starts in the year 2000 and is the

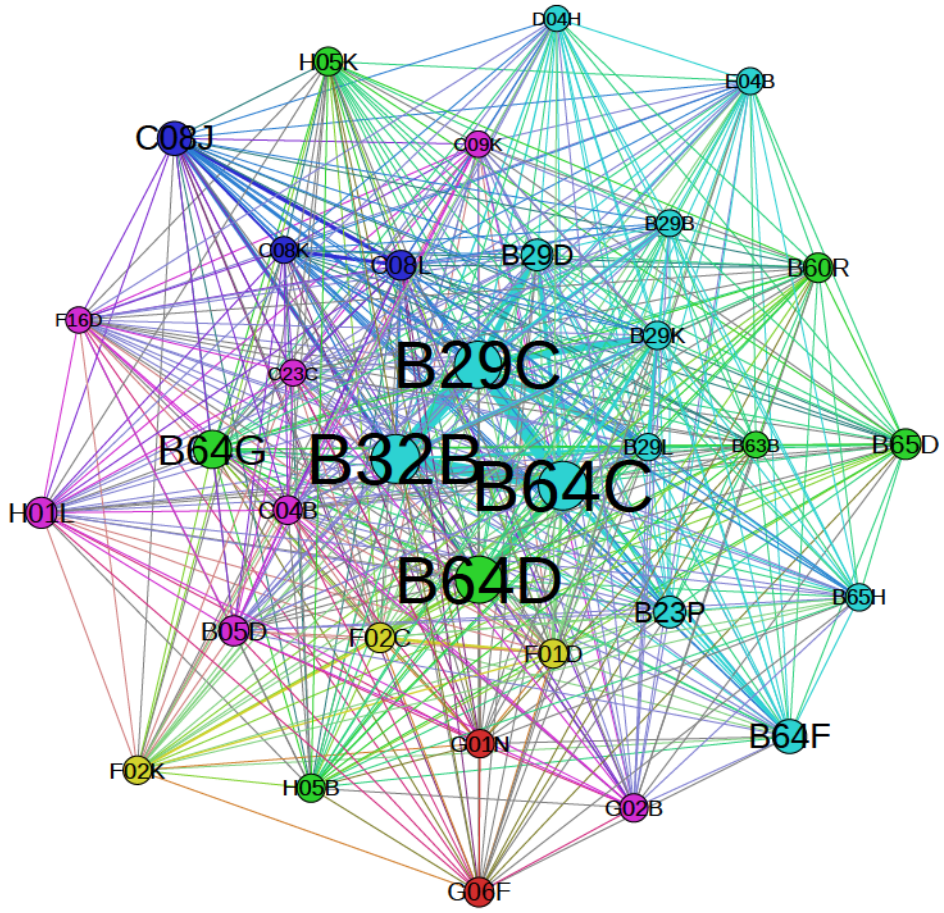


Figure 3.16: *The core interconnections of the knowledge network of SCM technologies in 2014*

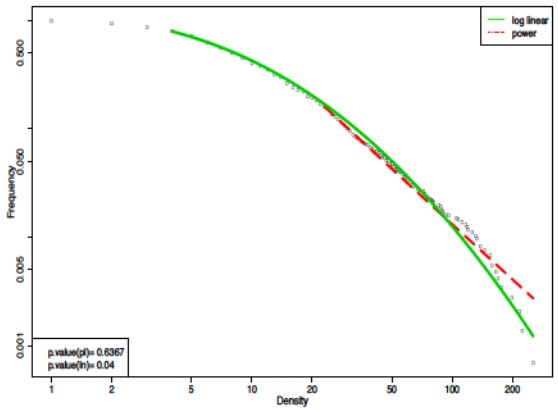


Figure 3.17: *Powerlaw and log-normal fit for the 7-digit network for 1985*

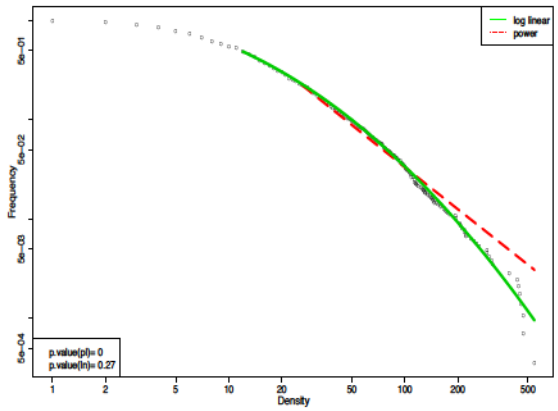


Figure 3.18: *Powerlaw and log-normal fit for the 7-digit network for 1995*

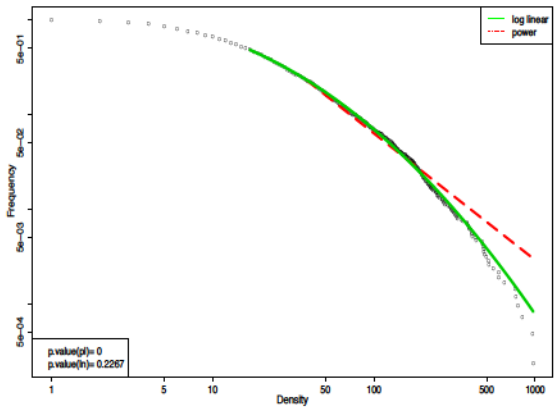


Figure 3.19: *Powerlaw and log-normal fit for the 7-digit network for 2010*

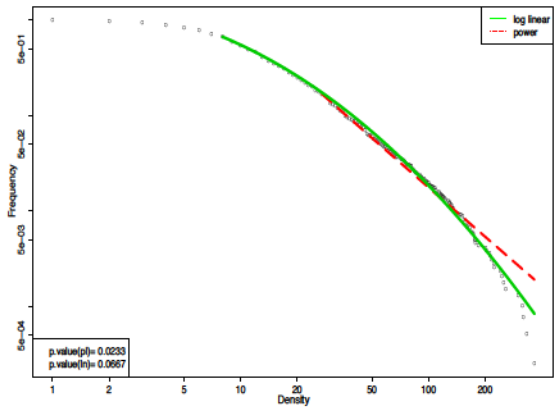


Figure 3.20: *Powerlaw fit for the 9-digit network for 1985*

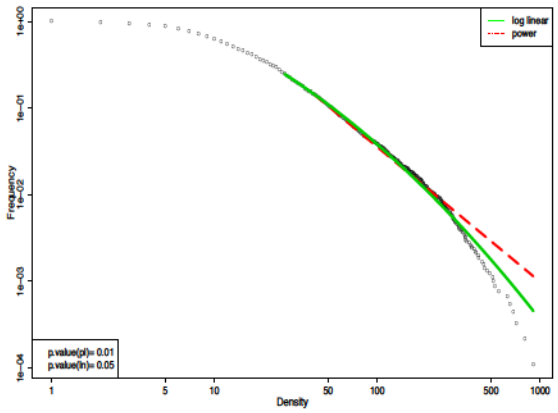


Figure 3.21: *Powerlaw fit for the 9-digit network for 1995*

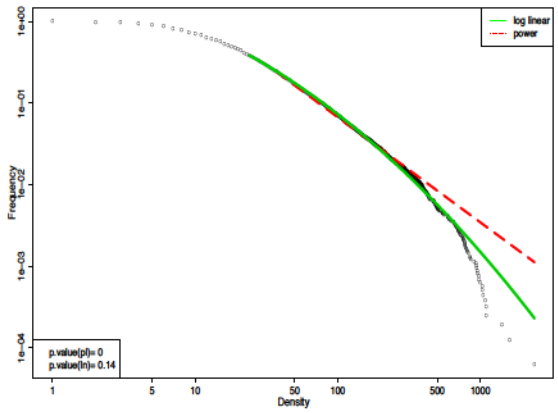


Figure 3.22: *Powerlaw fit for the 9-digit network for 2010*

result of existing actors in the network collaborating more extensively. New links are added between firms already in the network resulting in a decline in average distance and increase in clustering.

By connecting the patent network to the publication network we obtain the complete collaboration network. Firms that both deposit patents and publish in scientific journals will create connections between the two types of networks. The firms interconnecting the type network have a type of "gatekeeper" position. The gatekeepers are listed in table F.1 accompanied by the year of first appearance as a gatekeeper. We notice here that mostly the large companies that interconnect both networks rather than large research institutions.

The evolution of the complete network can be found in figure 3.27. For both networks we identify an inverted U-shape. We can hence distinguish two phases, a first phase staring in 1980 and ending around the year 2000, a second phase till the end of the period. During the first phase, identified as the research phase in the IPC network, the average distance in the network continuously increases. Mainly large multinational firms (e.g Honda, Taylormade golf, Rio Tinto, Saab, Astrium, Constellium, Daimler) enter the network during this phase. Since these actors have they own communities in the network the average distance increases during this period (the communities are interconnecting). When universities start to enter the network around the year 2000, the distance between the firms decreases. Universities typically take a central position and tend to have a large number of cooperations. In this sense they connect to many firms in different communities in the network reducing the distance. The appearance of research institutions also marks the start of the development phase of the technology. The fundamental technology has been developed by the firms in the previous stage. Collaborations to find applications for the technology are launched in the second phase. The IPC network already showed that the patents deposited during this phase relate to different developments around the fundamental technologies.

The data show then that the structure of the complete collaboration network converges towards a small world structure, this structure is reached around the year 2005. We can hence not validate hypothesis 4. The data does not show small world features as we expected. However, hypothesis 5 appears to be valid. When the IPC network changes from the research stage to the development stage we appear to observe a change in the structure of the collaboration as well. In the case of SCM the entrance of universities in the network

reduces the overall distance and causes the network to converge to a small world network. This means that once the technology is ready to be diffused, the structure of the network is optimal for knowledge diffusion.

The results show that the large assembler firms were the ones to develop the technology. The two largest firms, Boeing and Airbus, had diverging strategies when it came to the development of these technologies. Boeing collaborated with experts in the field of structural composite materials, while Airbus developed the technology with its historic partners. The next section will analyze these diverging strategies. The results in this section were discussed with engineers from Airbus. These discussions also helped in the interpretation of the observations.

3.5 Airbus Vs. Boeing: the impact of social proximity in link formation

We will now turn our attention to a specific part of the network. In the aerospace sector, two major competing actors are of interest: Airbus and Boeing. As we discussed before, these firms need to be able to learn all technologies related to SCM in order to include them into the production of the final product. Based on what we previously discussed, we can imagine two diverging strategies in terms of knowledge absorption for these firms. Either they decide to cooperate with firms from other sectors that have experience in the field or cooperation is based on social capital, i.e they pick firms based on whether or not they have previously cooperated. In this section we will show how these two strategies lead to different results in terms of innovative performance.

3.5.1 Network position

Boeing was present in the network since the 1980's while Airbus entered the market a couple of years later (this is why all graphs start in 1985). A first clue giving away the diverging strategies can be observed in Figure 3.31. When Boeing entered the market and until 1998, it has a clustering coefficient of 0. The clustering coefficient measures the number of collaborators of Boeing that work together. A clustering of 0 means that none

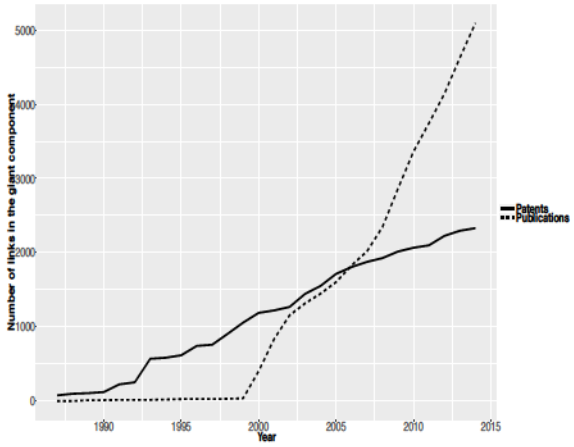


Figure 3.23: *Evolution of the number of links*

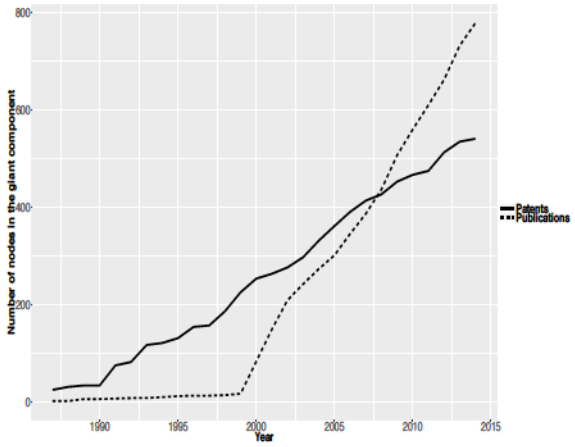


Figure 3.24: *Evolution of the number of nodes*

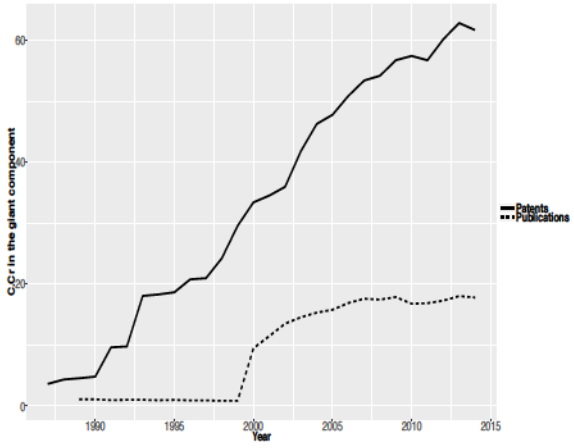


Figure 3.25: *Evolution of the adjusted clustering coefficient*

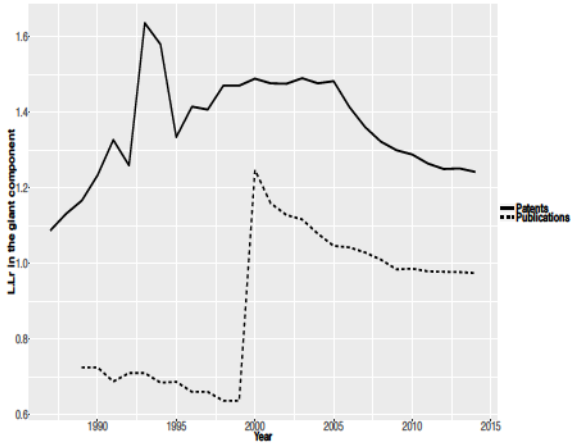


Figure 3.26: *Evolution of the average distance*

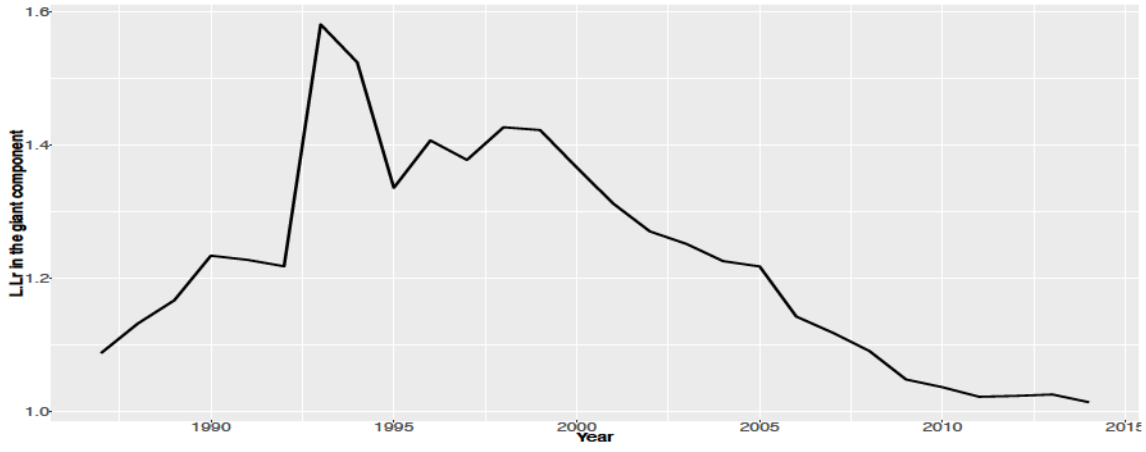


Figure 3.27: *Evolution of the adjusted average distance for the complete network*

Figure 3.28: *Dynamics of the collaboration network for SCM*

of the collaborators of Boeing have worked together. At the complete opposite we find Airbus which entered with a clustering of 1. Where Boeing took the risk of collaborating

with firms specialized in composite materials[1], Airbus chose to cooperate with a cluster of its historical partners. The strong links the firm has created in the aerospace sector have highly influenced its absorption strategy.

These observations are reinforced by the study of the citation network of both firms. Figure 3.31 shows the firms cited by Boeing and Airbus and the firms citing Boeing and Airbus. We observe here that the European company is largely influenced by the firms it has previously collaborated with while Boeing has a larger variety in its inspirations. At the center of this graph we find firms that inspired both firms. Above Airbus we find firms that inspired only Airbus, underneath Boeing we find firms that only inspired Boeing. The larger the arrow the higher the number of citations between two firms. We hence observe that Boeing has a large variety of inspirations with a low frequency while Airbus tends to cite more frequently the same firms. These firms are often previous collaborators.

The theory on preferential attachment Barabási and Albert (1999) suggests that firms might motivate their decision to cooperate with a specific firm if they have previously worked together. Firms who know each other have the advantage of cooperating more efficiently because they know how the other operates.

Boeing chose to identify specialists, with the risk that it would lose in efficiency during the cooperation because of a lack of social capital.

Towards the second phase of the evolution of the network we observe that both firms have very similar positions in terms of centrality. The betweenness centrality measures the relative position of a firm on all the paths connecting all other firms. This means that a high betweenness centrality is synonymous with a position through which many information flows may be captured. Figure 3.29 shows that both firms converge to a similar level of centrality as can be observed for the other indicators. This implies that both firms had an identical position for the absorbing of new technologies which is expected considering their position in their respective value chain.

The number of cooperations marks however a point of divergence. Airbus has accelerated the number of firms with which it collaborates from 1997 onwards. Boeing on the other hand has been much more conservative, ending with less than half the collaborations of Airbus.

We observe two different strategies resulting in nearly identical network positions. The

[1]We refer here to firms that have worked with composite materials in other sectors

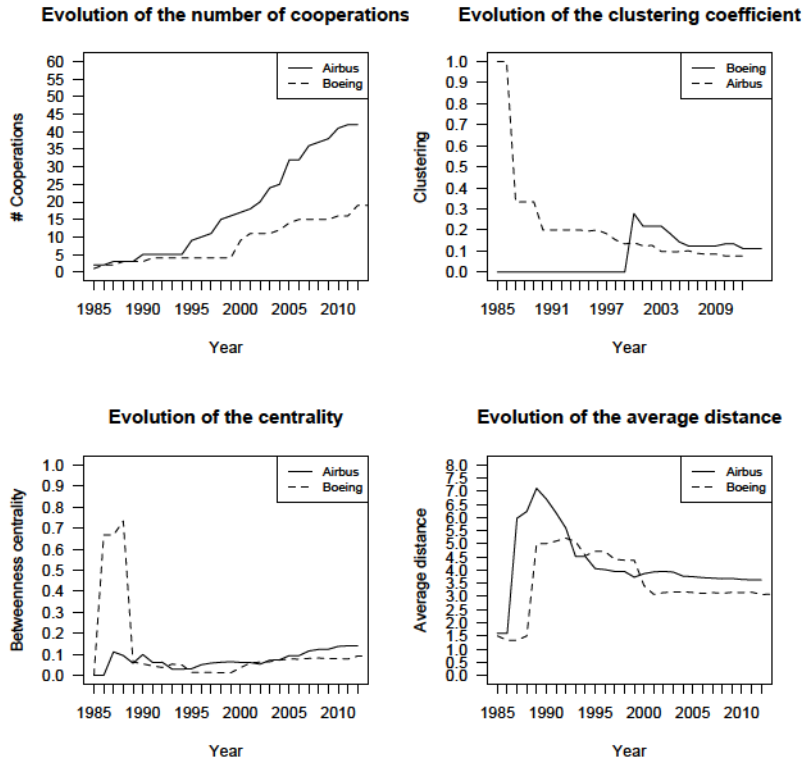


Figure 3.29: *Evolution of the network position of Airbus and Boeing*

resulting innovative performance is however not the same as we will show in the next subsection.

3.5.2 The race for innovation

By extracting IPC codes from Boeing and Airbus' patents we are able to track when firms deposit patents in specific IPCs that are at the core of SCMs technology. Two of these core IPCs are: B64C1 ("Fuselages, Wings, stabilizing surfaces, or the like"[2]) and B29C70 ("Shaping composites").

From the patents we create a network connecting IPC codes with Airbus and Boeing. If there is a link between Boeing and B64C1 that means that Boeing has deposited a patent using this code. The results are shown in figure 3.30, the thicker the link between the IPC and the firm the more deposits using the IPC were identified.

In the center of the graph we find IPC that were used by both firms, the most relevant IPC codes are found here. From a dynamic perspective we can observe when a firms first

[2]Titles given by WIPO in the international patent classification

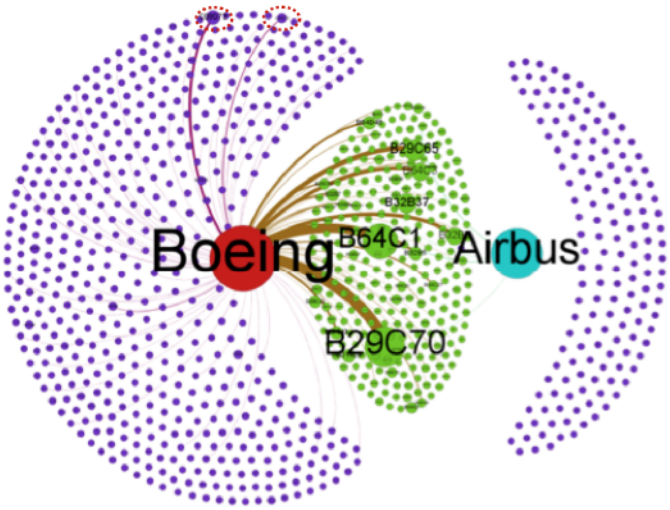


Figure 3.30: *IPC network of Boeing and Airbus*

deposits a patent in one of those relevant IPC codes. We hence created a dynamic network that allowed us to observe when a firm deposits its fist patent in a certain technology.
This representations can show how far a head (or far behind) a firm is compared to another. In our case we observed that Boeing deposited in the 2 IPC codes 10 years before Boeing did, showing clearly that Airbus has a technological lag compared to Boeing.
This lag can be explained by the previously identified strategy of Airbus, who decided to research the technology with historical partners. This decision was made even though the identified partners might not have been the most specialized firms in the sector.

Boeing's strategy paid off, it positioned itself as a gatekeeper between two sectors and it took the risk of collaborating with firms it has no connection with. Their knowledge absorption strategy was hence more efficient.

3.6 Conclusion and discussion

This chapter shows that the evolution of the IPC network can identify different stages of the technology life-cycle. This method can be used to identify at which stage of the life-cycle a technology is positioned. Firms can use this information in their business strategy when it comes to the identification of potential partners for example.

The IPC network has also shown that the initial technology was not developed by research institutions by the firms from the defense tier of the aerospace sector. Of course,

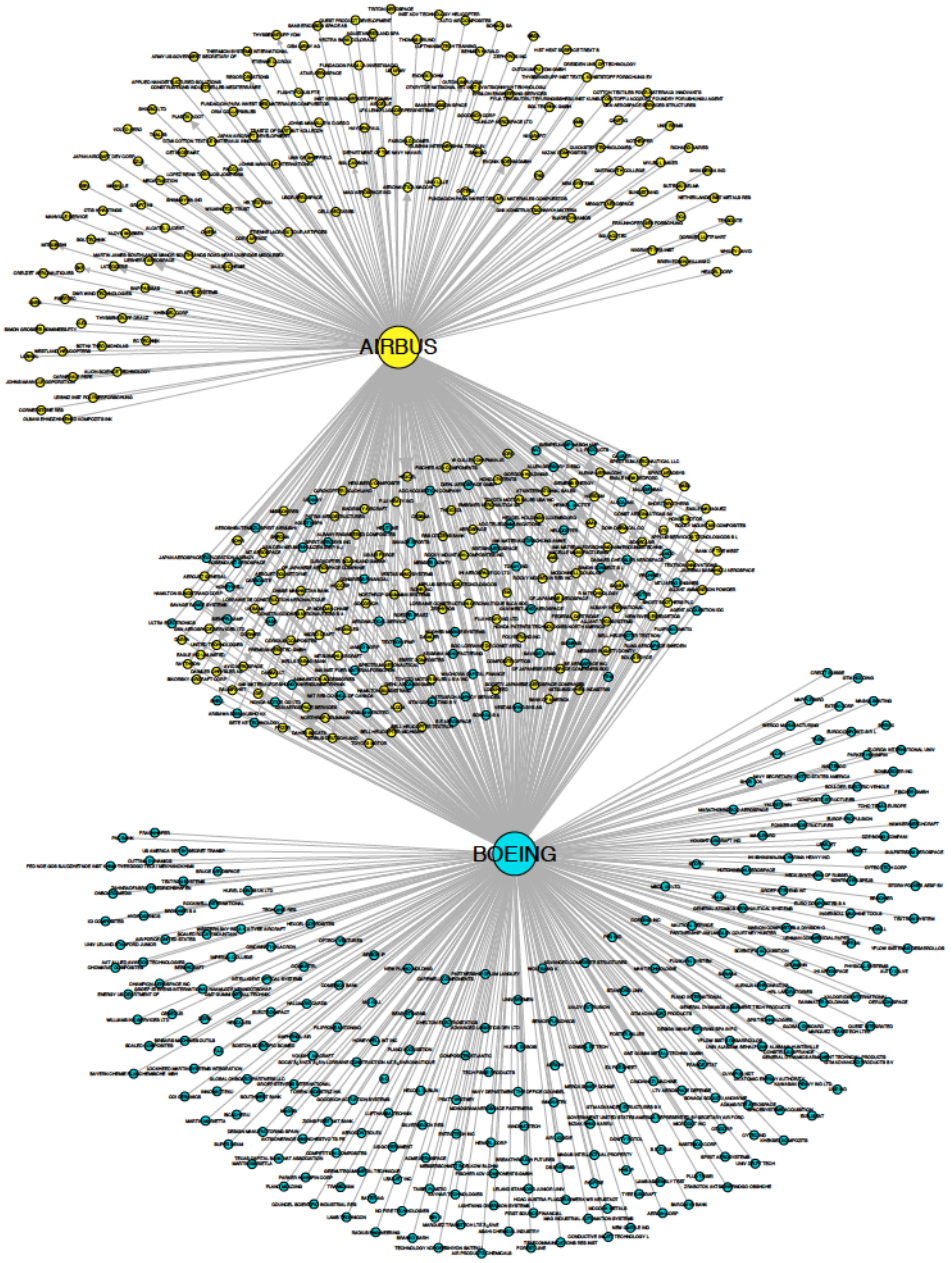


Figure 3.31: *Firms cited by Airbus and Boeing on their patents relative to the development of Structural Composite Materials in Aeronautics*

composite materials have existed before for other applications, but the results show that firms brought the technology to the aerospace sector. This has led to patents in chemistry deposited by large firms in the aerospace sector. The development phase is characterized by the appearance of research institutions who played a vital role in the development of applications for the technology. The evolution of the knowledge network presented the hypothesized characteristics, the first three hypotheses are hence validated. A core is formed during the research phase, while the development phase shows the appearance of a periphery around the core. In the collaboration network the evolution of the structure was less clear-cut than expected. In the early stages the average distance was too high to be a small world. The high average distance was due to a weak interconnection of the different clusters. These clusters started to regroup during the development phase reducing the average distance and resulting in the appearance of a small world structure, confirming hypothesis four. When comparing the knowledge and collaboration networks we can identify the year 2000 as point from which structures change. The switch in the knowledge network states a change in the type of technology deposited (as proven by the IPC codes on the deposited patents during that phase) while the change in the collaboration network shows the entrance of new firms and universities (as shown by both the patent depositors and the publication authors). There hence is a clear correlation between the evolution of the collaboration network and the knowledge network in the case of SCM technologies. So, when we analyse the evolution of the structure of innovators' collaboration networks, we should also take into account the development stage of the technology in its life cycle. This stage may be an important a structuring factor, as we have shown in this chapter.

Chapter 4

The evolution of the French Aerospace network

"A straight line may be the shortest distance between two points, but it is by no means the most interesting" – The third Doctor

4.1 Introduction

In this chapter I continue the analysis of the structural dynamics of innovation networks. While the previous chapter is focused on the level of the technology, I will now switch the focus to the level of the sector. Innovation processes behave differently according to their setting. In particular, as pointed out by (Pavitt, 1984; Hagedoorn and Narula, 1996), the sector is a defining factor in the innovation process. It would hence be interesting to study how the structure of an innovation network behaves according to the sector of analysis. This chapter will focus on the innovation network of the French Aerospace sector. I chose this sector for two reasons. First, it is a high technology sector that plays an important role in the French economy as well as the european economy. Second, the sector is organized in a particular manner, it is a value chain. This value chain has been been optimized by Airbus with its Power8 program. This particular type of sector should transpire into the structural dynamics of the collaboration network. The analysis of the structure of the network in this chapter will be pushed further than the method used in the previous chapter. Three levels of analysis will be presented, the global network level, the level of the clusters as well as a micro-level. The latter will be be accomplished using the ERG models presented in chapter 2. The aim of this analysis is to identify which factors incite firms to collaborate with one firm rather than another. The sector is a large supply chain build around the

European assembler Airbus. The latter has is a prime example of a modular firm in the sense that it has externalized most of its production to suppliers. The different parts of the aircraft are produced by different sections of the production chain. Each of which contain pivot firms (Frigant et al., 2006) that link the different parts of the aircraft together. In addition, since the year 2000, Airbus has been working on its "Power8" program, aiming at the optimization of its supply chain. Given these characteristics we would expect that the collaboration network of the French aerospace sector will closely resemble that of the production chain. Since the production chain is build up from a small number of highly connected firms (pivot firms) and a central assembler (Airbus) I propose the following hypothesis:

Hypothesis 1a: The structure of the collaboration network of the French aerospace sector is a core-periphery structure.

In order to better understand how this structure came to be the mechanisms that drive link creation between firms need to be identified. In other words, I want to know why did firm "i" collaborate with "j" rather than "k".
Technological proximity between firms is a requirement for cooperation. If firms are too similar, they work on the same technologies and hence would not want to collaborate. As the technological distance increases the complementarity of the knowledge bases of the firms increases. This results in an increase in the probability of observing a collaboration. This complementarity does however reach a point where technologies become too distant and the complementarity decreases. This results in turn in a decrease in the probability of cooperation. These statements induce the second hypothesis to test:

Hypothesis 1b: There is an inverted U-shape relation between the probability of a collaboration and the technological proximity of two firms.

In addition to technological proximity, social proximity is expected to play an important role when it comes to partner selection, especially since the "power8" program launched to streamline production. Reputation as well as similar work methods allow firms to work more efficiently by reducing frictions due to diverging methods. I hence propose the following hypothesis:

Hypothesis 1c: Collaborators of collaborators have a higher probability to collaborate than firms without a common connection.

Once the structure has been analyzed the focus switches to the link between the position of the firm in the network and its financial performance. As was stated earlier, knowledge flows through the network. According to the position of the firm inside the network, a firm can be exposed to more or less diverse knowledge flows, impacting its performance. Financial data on firms inside the network is used to measure the performance of the firm. I mobilize the Schumpeterian hypothesis that innovation is achieved by the recombination of ideas. This hypothesis implies that firm exposed to a large variety of ideas will have a high potential for innovation (Dosi, 2000; Cowan and Jonard, 2007). In other terms, the advancement on the inventive trajectory will be faster when the knowledge diversity available to the firm is stronger. When diversity is low firms risk decreasing returns to innovation. A variable called "neighborhood diversity" is used which computes for each year the number of technologies in the neighborhood of the firm. Each technology is considered to be an IPC code. The aim is to measure the diversity in the neighborhood, the IPCs of the focal firm are hence not included in the measure. This leads to the following hypothesis:

Hypothesis 2a: The technological diversity in the neighborhood of the firm has a positive impact on its performance.

Two theories claim the importance of clustering in a network. The two theories do however oppose each other when it comes to the sign of the impact. A first theory suggests that having collaborators work together results in a positive impact on innovation and performance. The cooperations allow for a better understanding of the functioning of each firm. This information will allow firms to better organize their innovative activities.The effect is enhanced when cooperations are repeated over time, the more they know about each other the more efficient the cooperation. The other theory however suggests that a social lock-in might occur when firms cooperate too often, they would rather work with people they know rather than take the risk of finding a partner that is not efficient. This may result in a reduction of the innovativeness of firms, by the means of a stagnation or even

reduction of the diversity of technologies. Instead of cooperating with a firm that masters new technologies they keep cooperating with firms that master the same technologies. This leads to the following hypothesis:

Hypothesis 2b: Clustering has a positive impact on the performance of the firm due a better mutual understanding of firms.

Notice that if this hypothesis is invalid then the theory on social lock-in would be valid. A network connects firms by creating paths between them. Knowledge flows between firms that are directly or indirectly connected. A firm with a position on many of these paths has access to more knowledge flows. This position is measured by the betweenness centrality coefficient which takes into account the position of a firm on path between other firms (Wasserman, 1994). The higher the centrality of the firm, the more it is on the crossroads of knowledge flows. The higher the centrality of the firm, the more it is able to benefit from diverse sources of knowledge.

The average distance gives a measure of the average distance a firm is removed from all other firms in the network. The closer it is to all other firms the more beneficial the knowledge flows should be. An argument against this idea is that if the distance is too low there is a high risk of redundancy of information and hence low distance should have a negative influence on the performance of the firm. I hence test the following hypothesis:

Hypothesis 2c: The more central the firm, the better the performance due to an increased access to knowledge flows.

The number of patents gives an indication of the innovative dynamism of the firm. The more patents are deposited by the surrounding firms the more knowledge they accumulated. The following hypothesis is hence tested:

Hypothesis 2d: The more patents in the neighborhood of the firm the stronger the knowledge spillovers to the focal firm.

Knowledge spillovers are only useful for a firm if she is able to absorb the knowledge it is exposed to. I use the number of technologies mastered by a firm as a proxy for the absorption capacity of the firm. The more technologies mastered by the firm the

easier it should be for the firm to learn new knowledge which should result in increased performance. This gives the final hypothesis.

Hypothesis 2e: The absorption capacity of the firm is positively related to its performance.

In this chapter I will first introduce the main assumptions for this sector, then the methods that will be used to determine the structure of the network, and the impact of the position of each firm in this structure on its performance.

4.2 Data

4.2.1 Patent data

Since our focus is on knowledge flows, data on collaborations that were initialized for the purpose of creating new technologies is required. For this purpose an innovation network is created using from patent data. Whenever two or more firms are present on the same patent a link is created between the firms. All patents were extracted from the Orbit database, the firm names in the dataset were treated by hand to remove any typos and text lost in translation.

I restricted the focus on Patents deposited in France by French companies in order to avoid any problems with data from different patent offices. For instance, the USPTO tends to cite more intensely than the other offices while the German firms make a heavier use of utility models. Restricting our dataset allows us to avoid biases in these aspects.

In order to select patents relative to airplane technologies a query was constructed using a combination of keywords and IPC codes. I found that using only keywords resulted in a heavy percentage of false positives while selecting patents according to NACE codes was too restrictive. The combinatory method allows us to focus on all the different technologies that make up an airplane. After all, an airplane is the perfect example of a multi-technology product (Prencipe, 1997).

Building such a query does require specific knowledge about the technologies inside an aircraft and their corresponding keywords and IPC codes. The query used here was

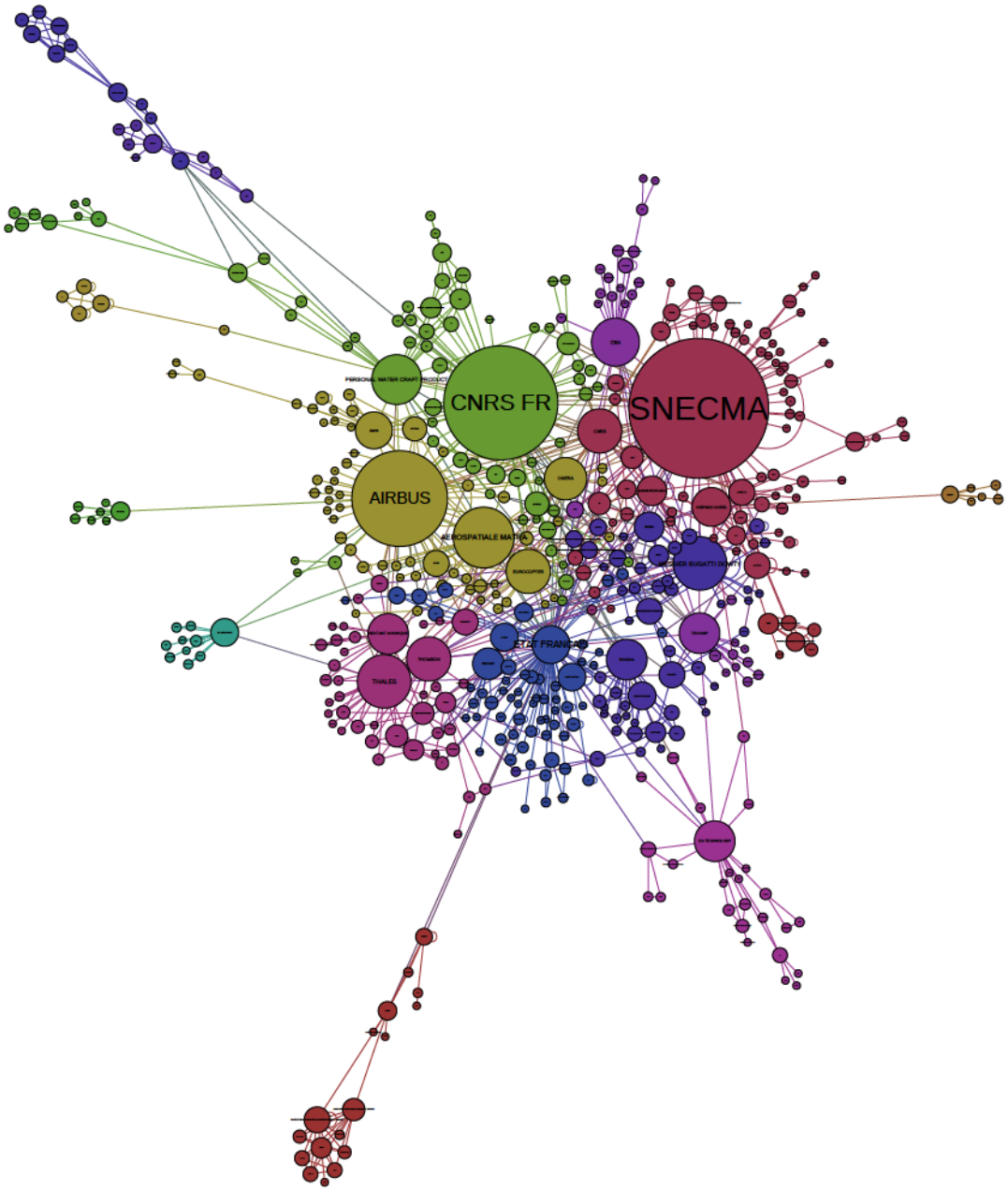
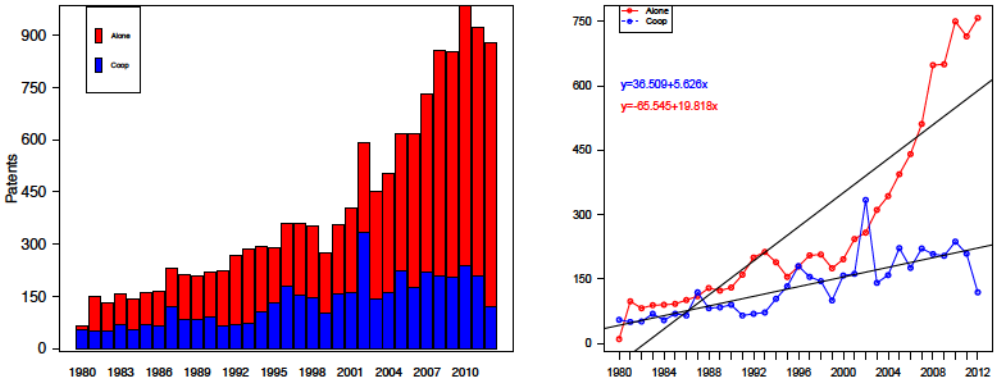


Figure 4.1: *The aerospace collaboration network as of 2014. Node size is proportional to the number of collaborations, colors correspond to structural clusters identified by a maximization of modularity.*



(a) Histogram of evolution of the number of patent deposits

(b) Trend in deposits

Figure 4.2: *Evolution of the number of patents and the corresponding trend. A distinction is made between the number of patents deposited alone (red) and the number of patents deposited by collaboration (blue)*

provided by the VIA-INNO platform[1] and is the result of repeated discussions between aircraft engineers and the platform to ensure viable results. The query resulted in a dataset of 11992 patents with a priority date between 1980 and 2013. 9544 (79.59%) patents were deposited by a single firm, 2448 (20.41%) patents were subject to a collaboration. From the 2448 patents that were identified 4369 cooperations between 1309 companies during the time period (1.78 cooperations on average per patent). Aggregation of these collaborations results in the network in figure 4.1.

Figure 4.2 shows the evolution of the number of patents deposited between 1980 and 2013. In figure 4.2(a) I distinguish between patents deposited by one firm and patents that are the result of a collaboration. Figure 4.2(b) shows a clear positive trend in both patenting and collaborative patenting in the aerospace sector. Similar observations have been identified in other sectors such as biotech and software by (Pyka and Scharnhorst, 2009; Gulati et al., 2011; Salavisa et al., 2012) . The trend for patenting alone (5.626) is however much higher than for cooperative patenting (19.82).

One can observe an important increase in the number of patents from the year 2000 onwards. This can be explained partially be the commercialization of the Airbus A380. A

[1]Plateforme d'intelligence économique labélisé centre d'investissement sociétale par l'initiative d'excellence de Bordeaux dans le cadre des investissements d'avenir de l'Etat Français (Website)

particular aspect of the aerospace sector is the fact that there are mass patent deposits after the commercial release of an airplane which might explain some of the variance in the dataset.

4.2.2 Financial data

The objective of the this section is to establish a link between financial performance and structural position. The structural position of the firm is important mainly because of knowledge flows. Innovations are achieved by the recombination of knowledge (Schumpeter, 1942). Since the knowledge stock inside a firm expands slowly and diversity decreases over time, external knowledge sources are important. The position of the firm inside the network defines the number and the diversity of knowledge sources to which the firms has access.

A panel data analysis will be presented to estimate the influence of the position of the firm on it's performance.

Financial data is hence required for the identified firms. From the sample of 1309 depositors all research institutions, financial institutions and government agencies need to be removed. 676 firms were identified in the dataset of 1309 firms. The financial performance of the firm will be measured by the Return On Assets (ROA) of the firms:

$$ROA_t = \frac{Net\ Income_t}{Total\ Assets_t} \tag{4.1}$$

The ROA seems the appropriate measure since the denominator of the ROA includes intellectual property and all capital mobilized for R&D activities. The data will be extracted from the Amadeus database. Since we have network data over 34 years it would be optimal to have 34 years of financial data. This was however not possible due Amadeus' policy. Firms are automatically deleted from the database once they have not transferred any data for 3 years. This means that firms that changed their names during the 34 year period are no longer in the database. Using DVDs from a previous version of Amadeus (between 2000 and 2007) it is possible to extract a relatively complete dataset over the years 2000 to 2012.

4.3 Methods

In order to check the hypotheses about the structure of the global network, methods are required. These methods are the same as those used in the previous chapter.

4.3.1 Core-periphery detection

The core-periphery structure is identified from the degree distribution of the network. A core-periphery network is defined a small number of densely connected firms and a large number of firms with a low number of links. Using the Cumulative Frequency Distribution derived from the degree distribution of the network one can fit a function to the data in order to test if the network has a core-periphery structure (see Appendix D for more details).

4.3.2 Small-World detection

In order to check if our network has a small world structure I follow a methodology presented by (Gulati et al., 2012). Small world structures are defined by a low average distance and a high clustering coefficient. The Clustering coefficient of a network is defines as the ratio of observed triangles in the network to the number of possible triangles. The average distance is simple the average number of links between any two nodes in the network.

Since nodes can be added each year I need to make sure that a decrease in clustering is the result of less firms connecting in triangles and not the simple result of an additional node that reduces the overall clustering coefficient. The coefficients are hence normalized and compared to a random network with an identical number of nodes and links.

The theory behind small worlds is that random networks have low clustering while empirical networks have higher clustering. The latter is the results of social / economic / geographic / ... motivations of the entities inside the network. As such, a network is a small world if its clustering coefficient is higher than that of a random graph of identical dimension (i.e same number of nodes and same number of links). This would hence imply that the graph is not random and that there are some underlying rules dictating the creation of ties in the network.

As for the average distance, it should be roughly identical to that of a random graph. I note Cr (Lr) the clustering coefficient (path length) of the random network and C (L) the clustering (path length) of the empirical data.
We hence need to observe $\frac{C}{Cr} >> 1$ and $\frac{L}{Lr} \approx 1$.
The evolution of the network was considered following two methods: using a 5-year sliding window and a method in which data was added year after year.

4.3.3 Exponential Random Graph Model

An Exponential Random Graph Model models the global structure of a network while allowing inference on the likelihood of a link between two nodes. It is basically a modified logistic regression, the models are modified in the sense that they do not require a hypothesis of independence between observations. For instance, if firm A is connected to B and C, there is a high probability that B knows C through its connection with A. A link between B and C has hence a higher probability than B connecting with a another, random, node. This implies that a link between two nodes depends upon the existing structure of the network. Regular logistic regressions are unable to account for these aspects since they require links to be independent upon each other. These levels of dependence are vital for the understanding of social and economic networks. The ERGM model to be estimated takes the form given in equation 4.2.

$$Pr(X = x \mid \theta) = P_\theta(x) = \frac{1}{k(\theta)} \cdot exp(\theta_1 \cdot z_1(x) + \theta_2 \cdot z_2(x) + \ldots + \theta_p \cdot z_p(x)) \quad (4.2)$$

Where X is the empirical observed network, x is the simulated network, θ a vector of parameters, z_i the different variables and $k(\theta)$ the normalizing constant. In short, the probability that the network generated by the model is identical to the observed network depends upon the given variables. If one consider that technological proximity has a role to play, it will be introduced as a variable. The model will then generate links while increasing (iteratively) the probability that nodes with higher proximity will connect. This is repeated a certain number of times. If, on average, the network generated is equal to the observed network then one can conclude that proximity plays a role the structuring of the network. For a more complete explanation of ERGM models see Chapter 2 of this thesis

(or Lusher et al. (2012)).

4.3.4 Measuring Technological proximity

Many measures of technological proximity exist, some are based on patent citations (Chang, 2012), (Marco and Rausser, 2008), (Mowery et al., 1998) while others use IPC codes (Jaffe, 1986), (Breschi et al., 2003). The idea is that the different technologies firms work on are not chosen at random, they co-exist because they have factors in common (Teece et al., 1994). This idea has led to different measures of technological proximity between firms, the most prominent was initiated by (Jaffe, 1986) further developed by (Breschi et al., 2003). Finer measures exist, see for instance (Bar and Leiponen, 2012) or (Bloom et al., 2013).

For the present chapter it is chosen to use an IPC based measure of technological proximity. A slightly different measure than the ones previously cited will be used, even though based on IPC codes. Our aim is to provide the likelihood of a cooperation based on the technologies mastered by firms. Therefore I assume that firms cooperate on technologies that are closely related in order to ensure proper incorporation of new technologies into an aircraft. As such having one technology in common is motive enough for two firms to cooperate. If one were to use one of the more common measures the prediction could be biased.

An IPC takes the following form: B64C1/18. Each part of the code (B, 64, C, 1,/18) indicates a practical classification. B stands for Performing operations and Transporting, B64 reduces the technologies to Aircraft, Aviation and Helicopters, B64C denotes Airplanes and Helicopters, B64C1 are Fuselages, wings etc. B64C1/14 are windows. The longer the code the more precise the technology. The full length of the IPC-codes is used in order to capture the largest amount of details of the technologies. When a firm deposits a patent one can deduce from the IPC codes what a firm has been working on and which technologies it masters. The measure of technological proximity is based on an analysis of IPC codes. The indicator of proximity computes the overlap in IPC codes between two companies. Table 4.1 shows two firms with 3 IPC codes. The numbers in the matrix correspond to the level of proximity. If both firms work on B they will have an overlap of one, if they both work on B64 the overlap is 2 and so-on. The proximity is maximal when firms deposit patent in the same 9 digit IPC codes. It takes the value of 0 when there are no elements in common.

		Firm B		
		B64C/19	B53D/01	C01F/03
Firm A	B64C/19	4	1	0
	B53D/01	1	3	0
	C01F/03	0	0	2

Table 4.1: *Illustration of the proximity measure used in the ERGM*

I defend the position that knowledge about one specific technology is enough to initiate a collaboration. The use of complete portfolios would induce a lot of noise in the data. In the end, firms cooperate often for a particular set of skills and not for all the skills used by a firm. A downside of this method is that the dataset is reduced to firms depositing both alone and by cooperation. One can only assume a firm masters a certain technology if it has deposited a patent alone. Cooperation data is then needed to create a network. Firms that only deposit by cooperation are hence excluded from the dataset.
A proximity matrix was computed for 176 firms and generated the network that connected them.

4.3.5 Variable lags for the panel regression

This study uses data from two different sources. The financial data from 2012 comes from the performance in the year 2012, the patent data from 2012 does however result from cooperations that took place any time before 2012. In order to perceive an effect of the cooperation on performance lags need to be included in the patent-related variables. How far back the lags should go depends entirely on the type of information, some have a faster influence on the performance than other do. In terms of lag we will consider that a cooperation is initiated three years before the priority date of the patent. This means that the transfer of some types of information may flow from that point on. The effects of the knowledge flow should be visible at about the date of priority of the patent. The effects of the production of the patented technology should be visible (if the technology is indeed put into production) at any point in time from $t-1$ on.
Structural variables: Firms are influenced by the knowledge held within the firm at the moment of collaboration. The diversity is hence lagged to $t-3$: firms connected by a patent in 2010 cooperated in 2007 and are hence influenced by the diversity in the firm

in the year 2007. However, since it takes time to absorb the knowledge and put it to use the impact on the ROA should be observed some time after the initialization of the cooperation, I will consider 3 years. Hence the variable Diversity is not lagged, the same is applied to the number of patents and the number of technologies. All the other variables are lagged at $t-3$ since the knowledge flows may influence the performance from the start of the cooperation on.

4.4 Results on the network structure

4.4.1 Cluster identification

The previously identified dataset leave us with over 4300 collaborations. The collaborations allow us to generate a network by creating a link between all firms that have deposited a patent together. The result is shown in figure 1. The bigger the size of the node the more collaborations the firm has. The coloring is the result of a community detection algorithm based on modularity. Modularity measures how well defined communities are inside a graph. Modularity gives a value between 0 and 1, the more the value tends towards 1 to more clearly defined the communities are (Newman and Girvan, 2004). For the result to be significant one expects a value of at least 0.6.

An algorithm introduced by (Blondel et al., 2008) was used to identify these communities using the open-source program Gephi (Bastian et al., 2009).

This community detection algorithm identifies communities inside a network purely based on the structural properties of the network. It starts by assigning each node with a community, it then selects a node at random and create a community with one of it's direct neighbors. The neighbor with whom it will create a community is the one that will maximize the modularity of the graph. This step is continued until maximum modularity is achieved. This method has the advantage of detecting automatically the number of communities (clusters) in the network while other methods ask the user for a fixed number of communities to be identified.

The results should however be handled with caution. The random component selects a node at random. It is possible that different results emerge if a different node is chosen at

the start of the algorithm. In fact, the sequence of choice of the nodes plays an important role in the detection of the communities. I hence ran the algorithm several times to make sure the same communities were detected on average.

The results are rather interesting given that the communities were clearly defined and easy to interpret. Different communities were identified around the following firms:

- Hispano Hurel: Nacelles
- Rhodia: Chemicals
- Thompson: Seating
- Messier Bugatti: Landing and braking.
- Pechiney Rhenalu: Structural elements (aluminium)
- Alcatel Lucent: Avionics and communication systems

These clusters suggest local technological development according to different parts included in the production of an aircraft. This allows us to understand the previously identified scale-free network structure. The large assemblers (Airbus, Snecma and Thales) and the CNRS have a large number of links connecting them with first order suppliers which in turn have their own clusters in which they are densely embedded.
This observation coincides with the industrial organization of the sector, which is indeed rather hierarchical. Airbus, at the center, designs the aircrafts while externalizing large portions of the production process to first order suppliers (Frigant et al., 2006). The latter will work with other, second order suppliers. As such there are not many competitors but competition is tough between the few (Niosi and Zhegu, 2005). The sector has undergone a significant restructuring in the 90' and the 2000's resulting in the specialization of some suppliers while others diversified their production to include other sectors (Frigant et al., 2006). In addition, the sector has high barriers to entry, mainly because of high level of knowledge required. The sector need an influx of cutting-edge technologies and hence close collaboration with fundamental research. The collaboration network that I observe here reflect these sectorial aspects: in a central position the CNRS (National Centre for Scientific Research) can be found providing an influx of fundamental science to the large

manufacturers and first order suppliers. While clusters exist around the first order suppliers connecting specialized and diversified suppliers. This results in a particular network structure that is made up from an interconnection of clusters. The overall structure of the network resembles a connected caveman structure (Watts, 1999) in which each specific part of the airplane is developed in it's own cluster. In terms of knowledge these firms need to collaborate with a large number of firms from different clusters in order to assemble an aircraft. While there is no need for direct knowledge flows between the landing and braking system and the nacelle manufacturer, Airbus needs knowledge on both technologies to assemble the final product.

The exception being that some firms connect all the clusters. Airbus has this central position since it needs to absorb knowledge from all clusters. Very little knowledge flows seem to exist between clusters, while there is a necessity for transfer intra-cluster.

Innovation in the aircraft industry is the result of an interplay of technology push and market pull (Dosi, 2000). On the one side aircraft manufacturers aim at making their aircrafts more cost efficient while there is a demand for governments to reduce noise and make planes more eco-friendly.

4.4.2 Structural Dynamics

In order to identify the structure of the network I will track the evolution of the network from 1980 onwards. This will allow us to have a clear vision of the structuring of the network.

Figure 4.5 reports the number of new firms that enter the network each year. The variance is explained by the previously discussed patenting behavior in the sector. The evolution of the number of nodes (figure 4.3) is computed using a sliding window of 5 years. This allows to keep track of the active firms in the network. This shows us that the network increases in size over the period with a decline during the last period (note that 2008 implies the frame 2008-2013). The decline can be explained by two factors. First, a small decline in the number of deposits in the last couple of years (figure 4.2(a)). Second, the decline in the number of firms might be explained by the "Power8" program launched by Airbus in order to optimize their production chain which resulted in a decrease in the number of suppliers. The evolution of the network was considered in two ways: using a

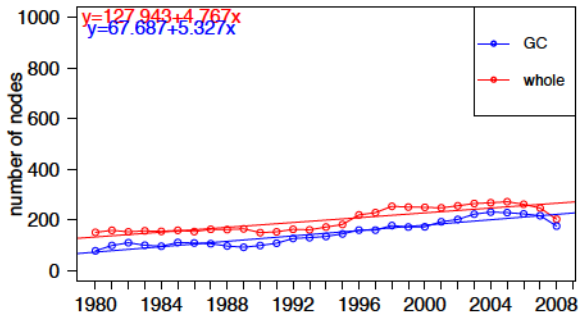


Figure 4.3: *Evolution of the number of nodes with a 5-year sliding window (i.e 1980 → 1980-1984). "GC" is the giant component of the network, "whole" the giant component with all the smaller components*

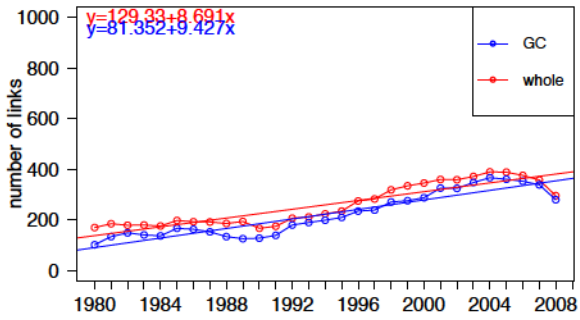


Figure 4.4: *Evolution of the number of links with a 5-year sliding window (i.e 1980 → 1980-1984). "GC" is the giant component of the network, "whole" the giant component with all the smaller components*

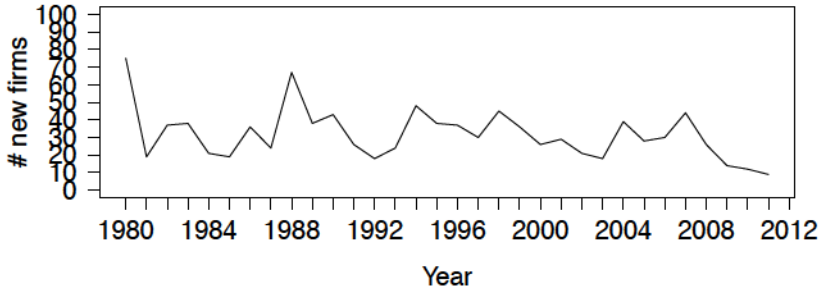


Figure 4.5: *Evolution of the number of new firms entering the collaboration network each year.*

5-year sliding window and a method in which data was added year after year. The results are reported in figure 4.6 and 4.7.

Figure 4.6 shows that the clustering coefficient trends strongly away from 1, indicating that the clustering observed in the networks increases faster than clustering in a random network of identical dimension. This is the case for both methods, showing that even when one removes firms that are no longer part of the network, the clustering stays higher than random. This high level of clustering is due to the different clusters that build the different parts of the airplane. These clusters are highly interconnected resulting in a high level of clustering. The power8 program which had the aim of optimizing the supply chain appears to have had a significant impact on the network, creating a decrease in the clustering coefficient that remained for a couple of years. The average distance shows a

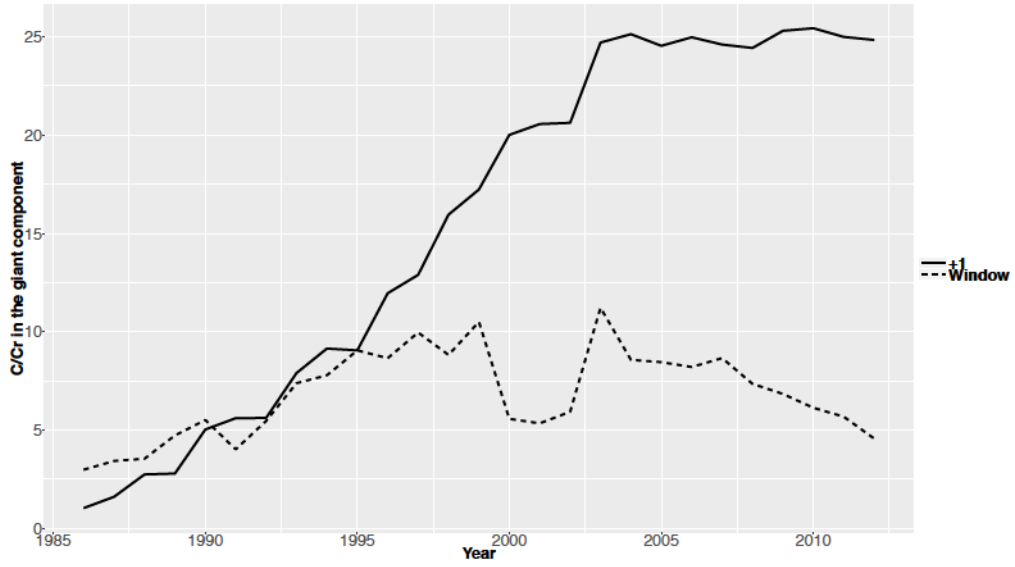


Figure 4.6: *Adjusted clustering coefficient*

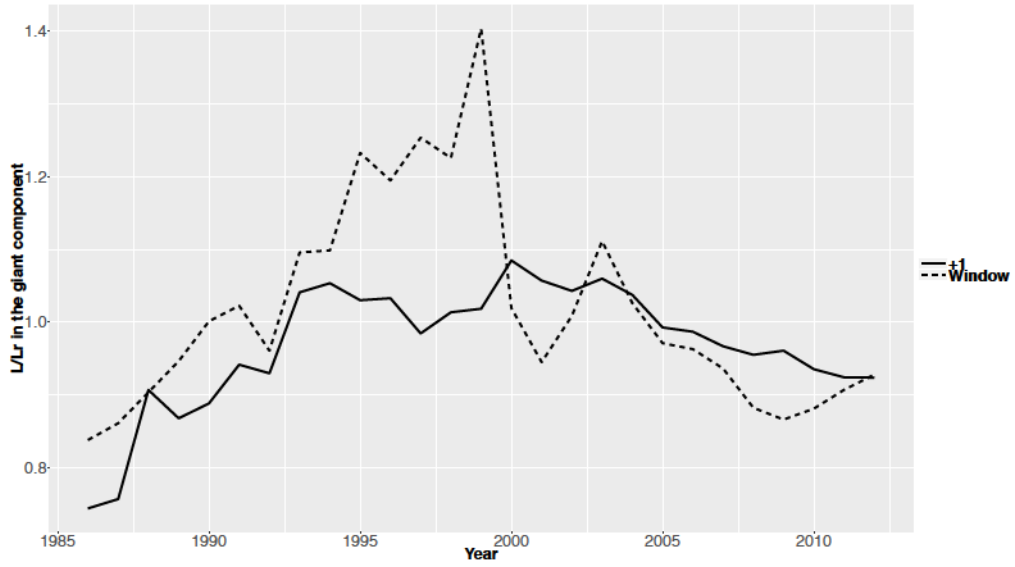


Figure 4.7: *Adjusted average distance*

similar decrease around this period, clearly showing the effects of the program. The 5-year window shows that the average distance of the network was too high (as compared to a random network) to be considered a small world. The different clusters in the network were not interconnected enough to be considered a small world. The drop in the year 2000 however, allows the network to reach the small world butter-zone. The +1 method shows that the network converges towards a small world early on and stays its course until the year 2007 where is converges towards the 5-years window. The network appears to have stabilized. I hence find converging conclusions from the results in Gulati et al. (2012) who

identifies an inverted U-shape in the small worldliness of the collaboration network. The structure of the network seems to be highly correlated with the structural specificities of the aerospace sector. Indeed, knowledge stays within the clusters since specific knowledge is developed inside each cluster. Knowledge flows between clusters through pivot firms interconnecting the clusters. Communication and knowledge flows are necessary between firms inside clusters since the parts developed by firms in clusters need to interact and need to be compatible. The most central firms hence benefit from the most knowledge flows since they have to assemble the different parts of the plane.

It can be concluded here that there is a high tendency for firms to cluster which confirms our previous observation that firms where organized in interconnected clusters. The structure also appears to stay relatively stable when it comes to these two indicators, especially in the time-laps network. In the 90' has started a radical change in the organization of the sector resulting in many suppliers exiting the sector which has as a consequence a lower number of collaborators. These collaborators collaborate more intensively resulting in a more stable structure towards the end of the period.

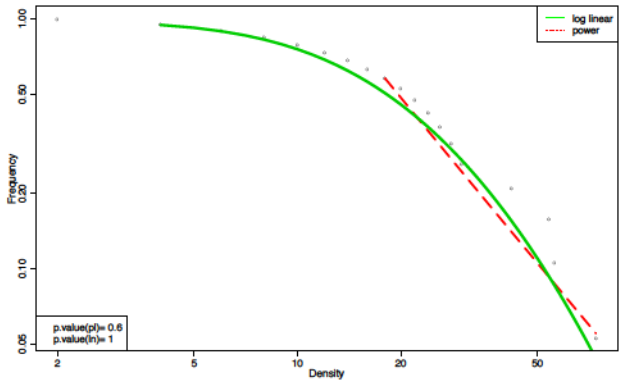


Figure 4.8: *Power-Law and log-normal fit for 1996 (window)*

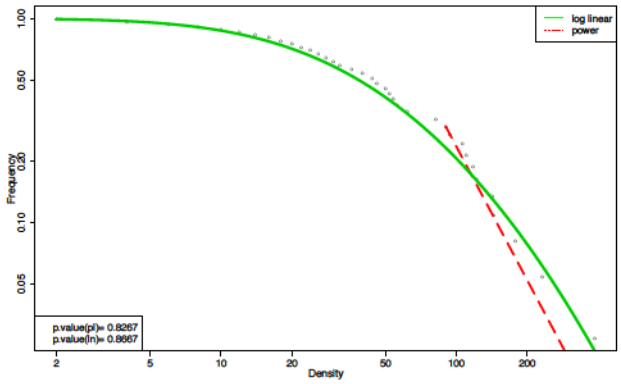


Figure 4.9: *Power-Law and log-normal fit for 1996 (+1)*

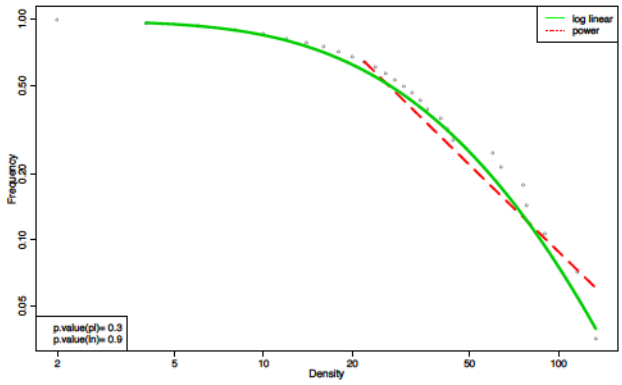


Figure 4.10: *Power-Law and log-normal fit for 2006 (window)*

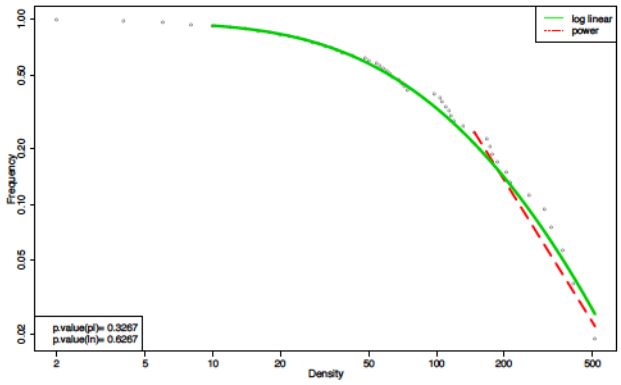


Figure 4.11: *Power-Law and log-normal fit for 2006 (+1)*

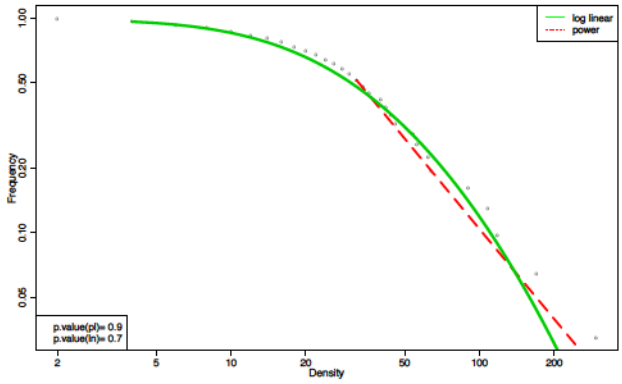


Figure 4.12: *Power-Law and log-normal fit for 2012 (window)*

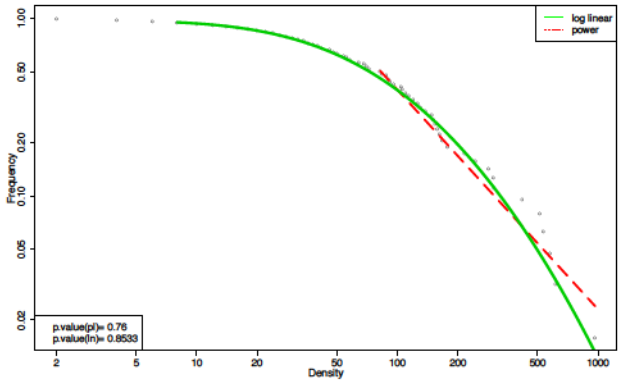


Figure 4.13: *Power-Law and log-normal fit for 2012 (+1)*

Quite interestingly, the network appears not only to have small world features but also core-periphery features. Figures 4.8 to 4.13 show the CFD of the network as well as the fitted functions. Recall that the null hypothesis (Data comes from a power-law

distribution) is rejected when the P.value is smaller than 5%. Even though the power-law is significant, it is only significant starting at a high density ($x_{min} \geq 20$). It cannot be concluded here that the network is scale-free. However, the log-normal fit is significant for both the window and +1 method. This implies that the degree distribution of the network follows a log-normal distribution stabilizing around $\mu = 3.44$ and $\sigma = 0.992$ (see table 4.2. The distribution shows that a large fraction of the nodes of the network have a relatively low density. At the same time, there is a low fraction of the nodes that have a relatively large density. The fraction of nodes with a low density is the periphery of the network. These are the firms inside the different clusters as can be seen in figure 4.1. The small number of firms with a higher density are the pivot firms, Airbus and the CNRS. The latter are connected to many firms inside the clusters to oversee the production of the different part they need to assemble. In addition they are connecting different clusters. The parts they create need to be compatible with other parts of the airplane. Interactions are hence required to ensure compatibility.

These elements result in core-periphery characteristics at the level of the global network structure. The network takes this structure from the early stages of the network until the end. The results in table 4.2 show the parameters of the adjusted law. The structure of the network stabilizes around the year 2005 for the +1 method, and a couple of years earlier for the window.

In conclusion then, the network has both small world and core-periphery characteristics. Similar results have been found in other types of networks by Guida and Maria (2007) and Requardt (2003). From these observations hypothesis 1a can be considered verified. In conclusion then, knowledge creation in the aerospace sectors is a localized phenomenon. Knowledge is generated in different clusters in which pivot firms assure the diffusion of this knowledge to the rest of the network.

4.4.3 Micro level motivations for collaboration

An ERGM model is used to determine the mechanisms that rule link creation. Table 4.3 shows the regression results, note that these coefficients cannot be interpreted as such. In order to compute the precise impact one needs to transform them into odds.

The results show that several factors explain the global structure of the network. It was hypothesized that technological proximity was a decisive factor in collaboration between

Year	Mean (window)	SD (window)	Mean (+1)	SD (+1)
1983	2.17	0.79	2.48	0.31
1984	2.55	0.56	2.28	0.56
1985	2.40	0.80	2.24	0.77
1986	2.55	0.58	2.26	0.80
1987	2.55	0.69	2.80	0.46
1988	2.30	0.90	2.47	0.87
1989	2.46	0.69	2.49	0.90
1990	2.28	0.89	2.52	0.89
1991	2.60	0.70	2.50	0.86
1992	2.65	0.67	2.78	0.76
1993	2.66	0.71	2.75	0.80
1994	2.50	0.87	2.82	0.82
1995	2.50	0.85	2.84	0.78
1996	2.73	0.57	2.95	0.80
1997	2.56	0.83	3.12	0.73
1998	2.65	0.67	3.18	0.73
1999	2.75	0.69	3.34	0.63
2000	2.68	0.78	3.10	0.84
2001	2.81	0.83	3.05	0.95
2002	2.81	0.86	3.27	0.75
2003	2.72	1.04	3.32	0.65
2004	2.63	1.04	3.30	0.74
2005	2.83	1.05	3.32	0.88
2006	2.86	1.03	3.33	0.90
2007	2.91	1.04	3.34	0.92
2008	2.86	1.07	3.28	1.04
2009	2.76	1.09	3.37	0.94
2010	-1.42	2.10	3.42	0.98
2011	2.43	1.13	3.44	0.99
2012	2.07	0.83	3.44	0.99

Table 4.2: *Evolution of the parameters of the fitted laws.*

firms in the aerospace sector. The models shows that this is indeed the case. Firms with a higher technological proximity have a tendency to work together. More precisely the odds of a link between firms that are technologically close is higher than the odds of a link between firms that are technologically far.

Moreover there appears to be an inverted U-shape to this relation as shows by the significance of the variable proximity2. This would imply that firms collaborate if they can learn from one another but if they are too close in terms of technology then the probability of a link deteriorates. Firms that are too close in terms of technologies can consider that the other firm has nothing to offer them and hence prefer collaborating with a firm that has

different technologies. Hypothesis 1b is hence verified.

The $altkstar$ parameter checks (and controls) for the core-periphery structure. Since the parameter is significant we see that the model has correctly identified the scale-free structure previously found.

Taking the $kstar2$ and $triangle$ parameter together allows for checking for triadic closure (Lusher et al., 2012) . Since both the parameters are significant I conclude that firms with a common node have a higher probability of connecting than firms with no common node. It hence seems that the trust that diffuses through the network as well as the increased performance due to common practices is a motivator for collaboration. Hypothesis 1c is verified.

Finally, co-citations are significant as well. Implying that firms that cite each-others patents will end up collaborating at some point in time.

4.5 Results on the impact of network position of the firm on performance

Two types of variables were included in this regression. Structural variables and technology variables. We have a panel of 1605 observations over a 10 year period. A standard linear panel regression to test the influence of the network on the performance of the firm is used. The previously discussed variables were included with the corresponding lags:

$$ROA_{t,t+1} = Clustering * density_{t-3} + Centrality_{t-3} + AverageDistance_{t-3} + Technologicaldiversity + Numberoftechnologies + Numberofpatents + Numberofcooperations$$

In a first regression only the variables relative to the position of the firm inside the network (model (1)) were used, a second regression includes only the technology variables (model (2)), the last model show the regression with both types of variables (model(3))

In order to assess which type of regression is adequate for the data several statistical tests were performed. The Lagrange Multiplier Test (Breusch-Pagan) showed that there is presence of panel effects in the data, simple OLS regressions are hence rejected.

I then checked for time fixed effects in the data, by adding a dummy variable for each year

	Dependent variable:		
	Network		
	(1)	(2)	(3)
edges	−7.267***	−1.121*	
	(0.228)	(0.626)	
kstar2	0.155***		
	(0.003)		
degree2		−1.336***	14.467***
		(0.255)	(2.970)
edgecov.citation			−20.934***
			(1.090)
triangle	3.428***	1.923***	1.726***
	(0.007)	(0.0001)	(0.0002)
gwesp	−0.439***		
	(0.166)		
gwesp.alpha	0.523		
	(0.385)		
edgecov.proximity2		1.565***	6.620***
		(0.271)	(0.345)
altkstar.1.6		−1.864***	
		(0.172)	
altkstar.1.7			−3.371***
			(0.086)
Akaike Inf. Crit.	578722	617651	9813
Bayesian Inf. Crit.	578760	617689	9851
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.3: *ERGM model results*

and compared the regression results with an F-test, the results show that no time-fixed effects have to be included in the model. A fixed, random and pooled model were then tested against each other, the fixed effects was retained as the best model. Since the data presented serial correlation and heteroscedasticity, I used robust estimates.

The results of the regression are shown in table 4.4. All variables have a significant impact on the ROA with the exception of the number of cooperations and the number of patents. The latter observation is rather to be expected. Not all patents have the same value and only a small portion of patents have an exploitable value. The number of cooperations shows that not all cooperations have a benefit in terms of knowledge flows. The number of collaborations being higher than the number of collaborators, it can be interpreted as the intensity of collaborations between firms, i.e how close firms are socially. The impact of social links is an order of magnitude lower than the impact of knowledge transfer by other objects and is difficult to capture.

The structural variables are all significant, showing that the position of the firm in the network does indeed have an impact on the performance of the firm. The adjusted clustering measure shows that firms with a higher clustering coefficient perform better. The collaboration of collaborators is hence a positive effect. The idea that working with people who already know each other seems to be validated.
In terms of knowledge absorption the central position of a firm is significant. The more central the firm is, the more knowledge it is able to absorb. The measure retained here is the betweenness centrality which measures the extend to which a firm is positioned on the a path between all the firms in the network. The higher the centrality the more favorable the position for knowledge absorption. The Average distance measures how far is firm is positioned from other firms, the further away the less knowledge the firm is exposed to. As such, the negative coefficient of this variable confirms the hypothesis that knowledge flows in the network have a decaying factor.

The technology related variables highlight the importance of technological diversity. Innovation literature puts forth the idea that innovations are achieved by the recombination of ideas. The diversity of technologies in the neighborhood of the firm should hence have a

	Dependent variable: Return on Assets		
	Network var.	Techno. var.	Combined
Adjusted clustering	0.646**		0.623*
	(0.313)		(0.322)
Centrality	0.890*		0.941*
	(0.513)		(0.501)
Average distance	−0.328**		−0.335***
	(0.128)		(0.127)
Technological diversity		0.002***	0.001***
		(0.0004)	(0.0005)
Number of technologies		−0.005***	−0.005***
		(0.001)	(0.001)
Number of patents		0.004	0.003
		(0.004)	(0.004)
Number of cooperations		0.001	−0.001
		(0.004)	(0.003)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4.4: *Panel regression results*

positive impact on the performance of the firm. The regression shows that this hypothesis is validated.

The final variable, the number of technologies mastered by the firm, has a negative impact. In our particular case, i.e the aerospace sector; the firms with the most technologies are suppliers with a specific position in the value chain. The regression show that specialized firms perform better than diversified firms, in a network. Specialized firms have to advantage of detaining valuable knowledge that can result in efficient innovations through collaboration. Diversified firms might be less interesting for cooperations and hence partner with less than optimal partners.

4.6 Conclusion

The production chain characteristic of the aerospace sector results in a network in which different clusters foster different technologies. These clusters are interconnected by a small number of large firms resulting in a Core-Periphery structure. The specificities of the aerospace sector play a vital role in the shaping of the collaboration network. The central position of Airbus in the networks ensure an interconnection of all different clusters. Knowledge is required to flow from each cluster this central firm. Knowledge is created locally in this network and diffuses through the pivot firms to the assembler. The Power8 program instigated by Airbus in the early 2000's had for main objective to streamline the production chain, and this appears to have had as a result a small world structure in the collaboration network.

On a micro-level this chapter has shown that technological proximity explains collaborations between firms but that this behavior follows an inverted U-shape. There is hence a butter-zone for the level of proximity that leads to collaboration.

The analysis of the performance of the firm tends to indicate that a central position in the network goes hand in hand with better performance for the firm. This is explained by the access to knowledge flows by firms with a high centrality and a low average distance. The choice of partner is proven to be important for two reasons, the clustering of the firm and the specialization of the firm. If the partner evolves in an environment in which collaborators of collaborators collaborate, this will have a positive impact on it's performance. If the firm choses a specialized firm to innovate with this will also have a positive impact on the performance of the focal firm.

Chapter 5

The evolution of the French Biotech network

"Never ignore coincidence. Unless, of course, you're busy. In which case, always ignore coincidence"

–The eleventh Doctor.

Introduction

The biotech sector is a dynamic and highly concurrential high-tech sector. While the aerospace sector is defined by a value chain with low competition on the national level, the biotech sector is defined by high competition and no value chain structure. The sector is defined by the importance of patenting, on which firms rely to protect their R&D efforts (Powell et al., 1996). Few other sectors rely as much on patents to secure returns from their investment as the biotech sector (Inventions, 2002). Due to this high level of patenting there are concerns that innovation has been slown down, especially since many of the research is fundamental research. The public research institutions that were the first concerned by this problem have since recognized the need for patents to secure returns. The marketing of innovations in this sector are under scrutiny of both national and international regulations. The health-risks that might be present in some of the products must be tested before a product is allowed to be marketed. The sector for biotechnologies is composed of four research directions referred to by colors. White biotechnology is the application of biotechnology for the processing and production of chemicals, materials and energy. Enzymes and micro-organisms are used to make products for other sectors (e.g. food, textiles). Blue biotechnologies are related to maritime research, exploiting

organisms living in the sea for the purpose of identifying new enzymes for example. Green biotechnologies relates to plants, searching for the developments of new plants able to survive on saline ground for example. The final color is red for medicinal biotechnology, based on the analysis of ADN. The collaboration network of biotech sectors has been studied in different countries by the means of different data sources; patents, polls, meritcati and biosource (Buchmann and Pyka, 2013; Kogut, 2000; Gay and Dousset, 2005; Quintana-García and Benavides-Velasco, 2008; Van der Valk et al., 2009). The Biotech sector has features that are on the opposite of those from the aerospace sector. There is no particular reason for firms in the different segments to collaborate since they have different domains of application. I expect to identify a clusters that correspond to the different tiers of biotech research. There should not be a central actor connecting the different clusters and hence i do not expect a small world structure. For these reasons, I think that a contrasting of this sector with the previous one should bring a better understanding of the dynamics of technologies and collaboration network. I propose the following hypotheses.

Hypothesis 1a: The network structure of the Biotech sector in France has a core-periphery structure. In highly competitive sectors such a biotech, firms are hesitant to share data and knowledge. Collaborations emerge for the purpose of combining knowledge more than risk-sharing (Powell et al., 1996).

Hypothesis 1b: There should be an inverted U-shape relationship between the probability to collaborate and technological proximity of firms.

Referrals and repeated interactions are hence of paramount importance:

Hypothesis 1c: The clusters in the collaboration network are highly clustered and well defined. Triadic closure should have a significant role to play in the formation of the network.

The financial analysis will seek a link between the position of the firm and its performance. Since the hypotheses are the same as those of the previous chapter I will simply recall the hypotheses here. The theoretical arguments are identical to those of the previous chapter.

Hypothesis 2a: The technological diversity in the neighborhood of the firm has a positive impact on its performance.

Hypothesis 2b: Clustering has a positive impact on the performance of the firm due a better mutual understanding of firms.

Hypothesis 2c: The more central the firm, the better the performance due to an increased access to knowledge flows.

Hypothesis 2d: The more patents in the neighborhood of the firm the stronger the knowledge spillovers to the focal firm.

Hypothesis 2e: The absorption capacity of the firm is positively related to its performance.

The following section will present the methods used for the identification of the relevant dataset of firms as well as their patents. Following this section, the network analysis will be presented including methods that were not used in the previous chapters (overlapping communities and their dynamics). With the structure analyzed, an ERGM model is presented that will identify different link creation mechanisms. The final section aims at identifying a link between the position of firms in the collaboration network and their financial performance. The final section concludes.

5.1 Data

For the aerospace sector a complex query was constructed using a combination of keywords and IPC codes. Since the biotech sector is rapidly evolving, the keywords, in the form of molecules, evolve rapidly. In an attempt to identify relevant key-words I extracted the text of patents deposited by firms that declared working in the biotech sector. The aim

is the identification of the largest and most accurate dataset possible. By extracting the text and creating a network of the words it is possible to identify relevant keywords. In order to make this network easy to use I removed all irrelevant information by computing the minimum spanning tree of the network. A minimum spanning tree keeps only the links that are part of the shortest paths between nodes. This results in the removal of a large number of links from the network reducing the clustering to 0. This makes the network easier to analyse without losing any vital information. The result is a tree-like network as shown in 5.1. The size of the nodes as well as the text is a function of the frequency of appearance. The colors are the result of a community detection algorithm that will be discussed later in this chapter. The keywords that emerge, as shown in Figure 5.1 are too generic and result in a high number of false positives. The tree shows that even the most central concepts are too general to be considered as keywords. A different methodology than the one applied for the aerospace sector had to be found.

The methodology used to identify relevant firms in the french biotech sector was to use the NACE classification[1]. Associations for firms in the biotech sector exist. I used a list of these firms and identified their NACE classification codes. Using the identified codes I proceeded to an extraction of all firms with these codes from the AMADEUS database. These names were introduced in the Orbit database for patents and all patents for the identified firms were extracted over a 25 year period from 1980 to 2015. I then extracted the co-assignees of these patents to double check if any firms were missing. After manual cleaning of the firm names on the identified patents I was able to identify 2061 patents deposited by french firms in France. Figure 5.2 shows the network as of 2014.

5.2 Methodology

5.2.1 Global network structure identification

I follow here the same methodology as in Chapters 3 and 4, the average distance of the network and the clustering which will be normalized by dividing them by the expected values from a random network of the same size (i.e same number of links and same number

[1]Statistical classification of economic activities in the European Community, comes from the French name: Nomenclature statistique des activités économiques dans la Communauté européenne

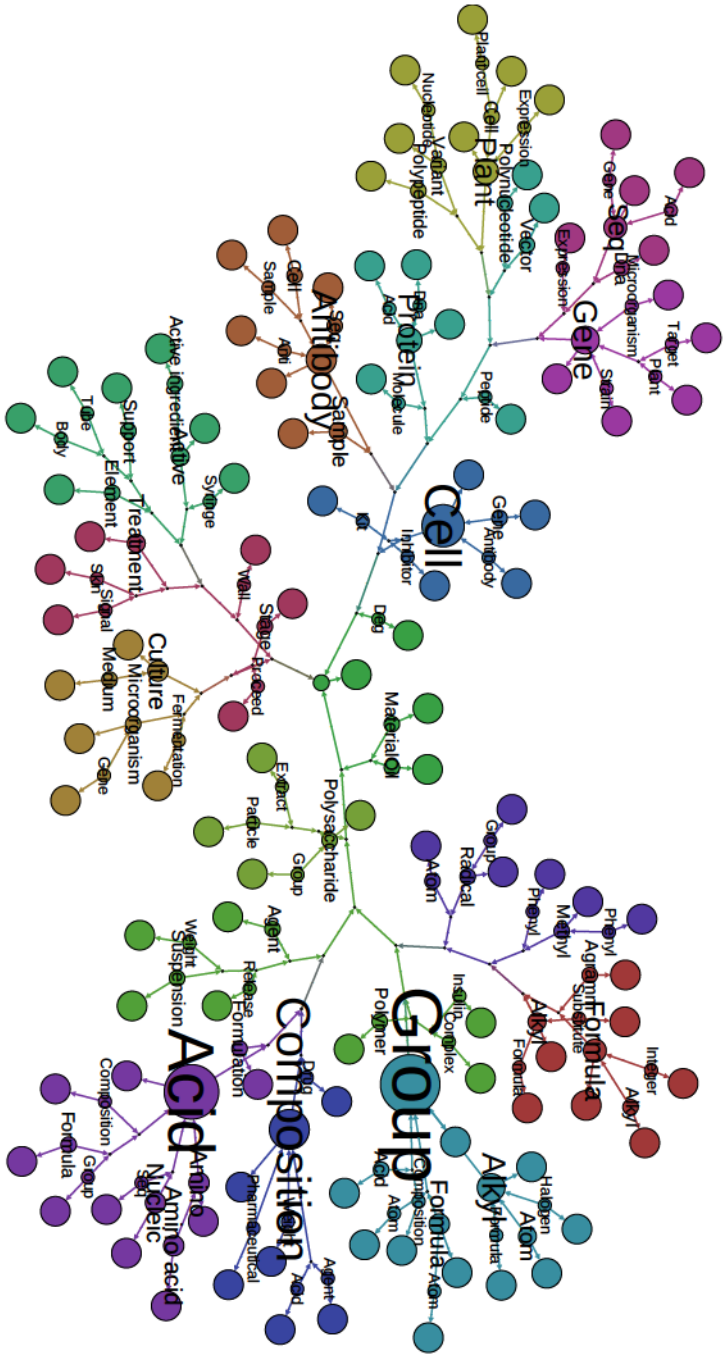


Figure 5.1: *Tree of concepts in the biotech patents. This is the minimum spanning tree of the concept network. Links are created between words in the summary of patents. The structure then highlights the central concepts and the specific concepts they are connected to.*

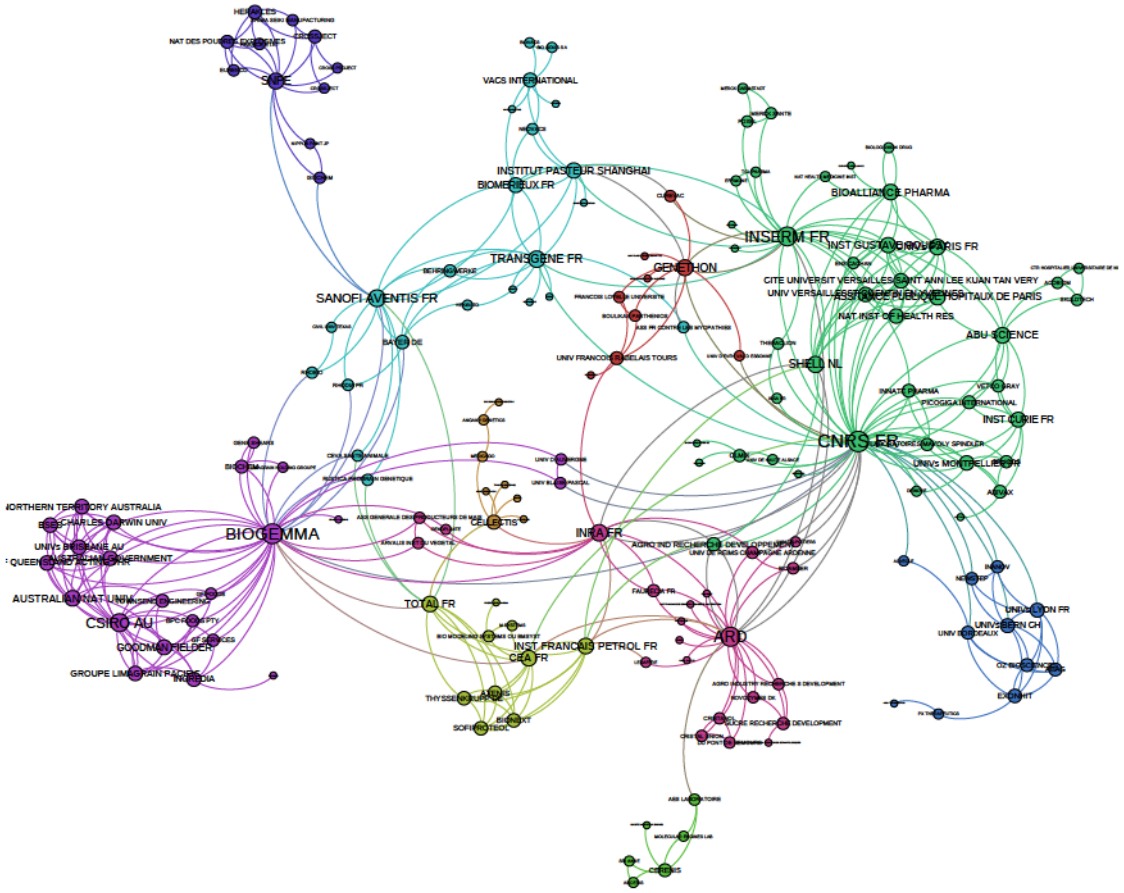


Figure 5.2: *The French biotech sector as of 2014*

of nodes). For the identification of the core-periphery structure the same method is used as well. The identification of communities is expanded by the use of a method for the identification of overlapping communities.

Community detection is performed by the means of Modularity maximization. Modularity measures how well a network can be divided in interconnected clusters. It maximizes the number of links inside clusters while minimizing the number of links between clusters. To identify a cluster the algorithm compares the edge density in each cluster with the edge density of the cluster in a randomized version of the graph (links are cut and reassigned at random).

More precisely if once considers a network that has been divided into communities one can generate a matrix in which each element a_{ij} represents the fraction of the nodes that are both in community i and community j. In this matrix the diagonal then represents the edges that appear in the same community. The trace of the matrix should hence give a measure of the quality of the repartition of the nodes into communities. But since the trace of the matrix can be maximized if there is only one identified community the measure needs to be more precise. If the trace is compared to a random network, i.e a network with the same communities but random connections the measure becomes more precise. This measure is called modularity which aims at maximizing the number of links inside a community while minimizing the number of links between communities.

A dendrogram is produced that starts with all the nodes in the network and ends with one final cluster. In between the nodes are regrouped in clusters which gives a value of the modularity statistic: Q. The dendrogram is cut at the value that maximizes Q as can be seen in figure 5.3. The blue line on the right shows the value of Q for each cut of the dendrogram. This automatically gives us a number of clusters. This is an advantage when compared to other methods that ask for a manual input of the number of cluster one wishes to find.

The coloring in Figure 5.2 identifies the different clusters using this method. The modularity value (Q) of this network is 0.783 which implies that the communities are well defined.

Following the idea that nodes can also be part of different clusters I use the R package "Linkcomm" (Kalinka and Tomancak, 2011) to identify overlapping communities. The latter uses algorithms proposed by (Ahn et al., 2010) based on the Jaccard coefficient for the

identification of similarities between links. A dendrogram is build from the identification of clustered links and cut at a level that maximizes partition density.

5.2.2 Financial analysis methodology

Network indicators were computed according to the addition method. In other words networks indicators are computed for 1980-1981 then 1980-1982, 1980-1983 and so on. The position of the firm in the network in the years 2000 is then the aggregation of all patents deposited since 1980. All structural variables as well as technological variables used in this regression will contain this historical dimension. Unfortunately the financial data is only available from the year 2000 onwards. The model can only cover the years 2000 to 2013.

The Return On Assets (ROA) was used an indicator of the performance of the firm and was hence the endogenous variable.

Since all firms are in the same sector I will not control for sector (unlike the aerospace sector), however, a control for years is used. The growth rate of the ROA is computed in order to reduce de impact of the size of the firms. The lags are the same as those for the aerospace sector (see section 4.3.5).

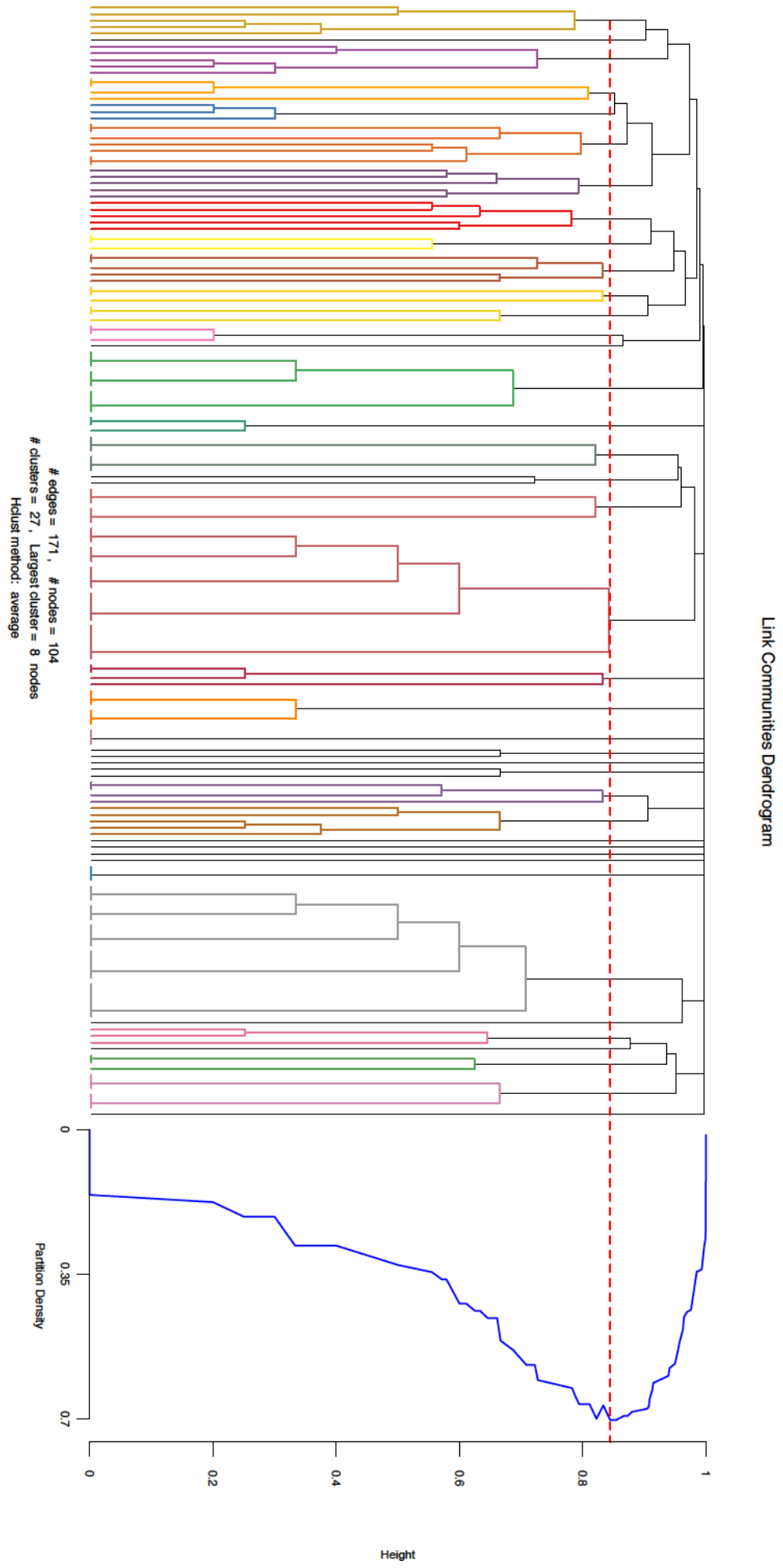
The estimated model is:

$$ROA_{t,t+1} \sim Clustering * density_{t-3} + Centrality_{t-3} + AverageDistance_{t-3} + Technologicaldiversity + Numberoftechnologies + Numberofpatents + Numberofcooperations$$

5.3 Results of the network analysis

5.3.1 Structure identification

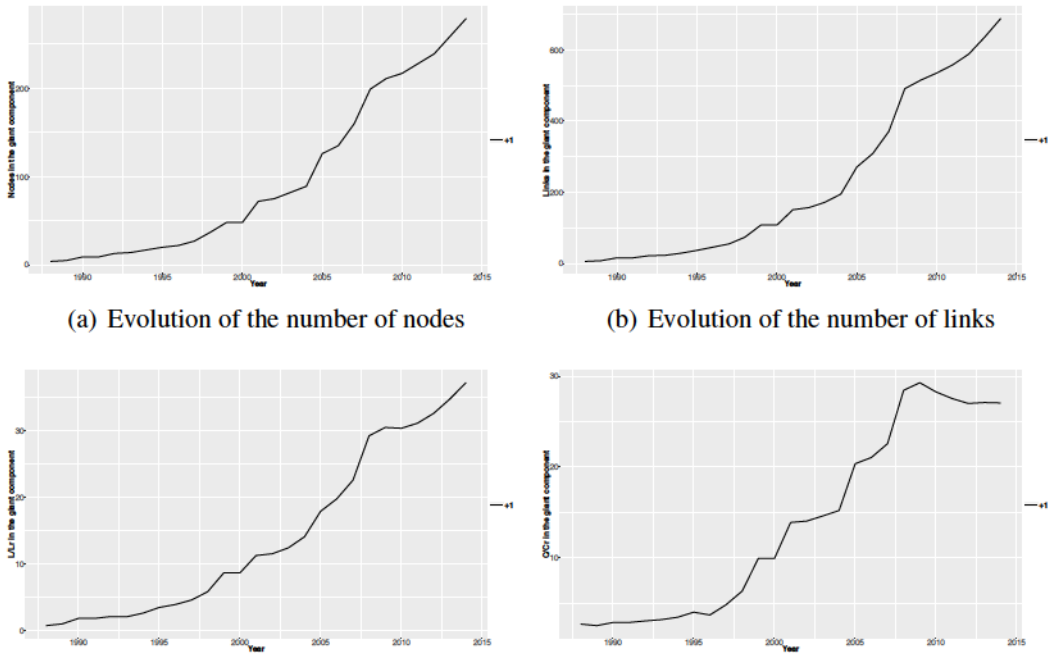
Figure 5.4 shows the evolution of the network indicators. The number of links as well as the number of nodes increases steadily over the period of the analysis, highlighting the need for collaborations in the sector. Figure 5.10(d) shows that the average clustering coefficient is thirty times higher than that of a random network. The network is presents a high level of clustering as compared to a random network. In addition, figure 5.10(c) shows that the average distance between firms is high. This shows that there are few links between the clusters. Firms have a tendency to collaborate with other firms inside their


Link Communities Dendrogram
edges = 171, # nodes = 104
clusters = 27, Largest cluster = 8 nodes
Hclust method: average
Partition Density
0
0.35
0.7
0
0.2
0.4
0.6
0.8
1
Height

own cluster. Knowledge is hence required to travel a long distance to reach all firms in the network. However, the small number of connections between the different clusters shows that there is only limites knowledge flows between communities. It would appear then, that knowledge does not need to travel through the whole network. The aerospace sectors has one common objective which is the construction of an airplane. The assemblers make the link between the different clusters reducing the average distance in the network. In the present case this link is absent and the average distance stays too high for the network to be considered a small world network. The question remains then to know if the network has a core-periphery structure. We hence turn to figure 5.5. The latter shows the cumulative degree distributions of the network at different points in time. The scale-free fit is not significant, neither is the log-normal fit. The end tail of the distribution is too irregular, no clear core appears here. There is a significant drop in frequency around a density of twenty in the last stage of the network but these nodes do not form a core. Rather, they are the highly connected firms in the different clusters. In other words, we have highly connected nodes, but they are not interconnected between themselves. The network structure does hence not present a core-periphery structure. Even though the network does not present the characteristics of any well known global structure, the structure is not random. Clearly, the results of the small world analysis show that clustering exceeds that of a random graph, and so does the average distance. In order to better understand these observations we turn to community detection.

5.3.2 Community identification

Figure 5.2 shows the network of the biotech sector in France in 2013. Nodes with identical color are part of the same community as identified by modularity maximization. The size of each node is proportional (non-linearly) to its density. The different communities appear to be well defined in the network. Each community is characterized by one or a couple of large firms. More importantly the different clusters are defined by different types of biotechnologies. Sanofi-Aventis, Transgène, Genethon and Biomerieux are firms that can be classified in red biotechnology research (medecine). These firms are connected close together in the network. ARD can be classified in white biotechnology, having its own cluster close to the public research institutions. This cluster also connects with Total. Total is a big company and does not focus on one specific type of biotechnology,



(a) Evolution of the number of nodes

(b) Evolution of the number of links

(c) Evolution of the adjusted average distance coefficient

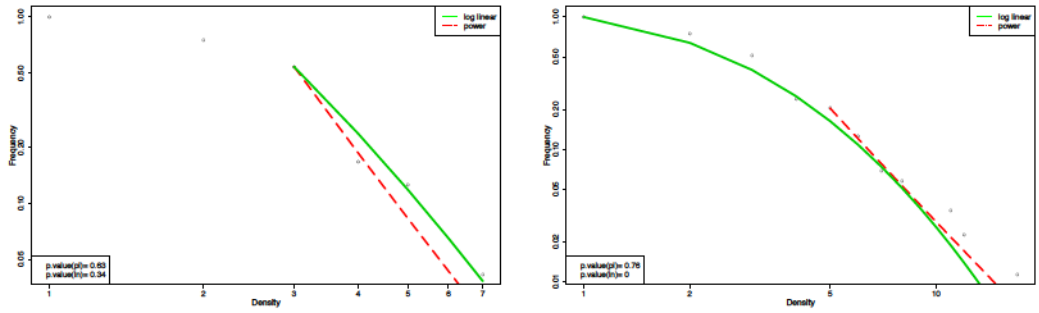
(d) Evolution of the adjusted clustering coefficient

Figure 5.4: *Structural dynamics of the French Biotech collaboration network between 1986 and 2013.*

Total focuses both on green and white biotechnologies[2] which explains its position in the network connecting the white cluster from ARD and the green cluster of Biogemma. Overall the network is a representation of the different types of biotech. Different clusters are interconnected by firms working on several types of biotechnologies. Since firms can work on different types they should be included in both clusters, in theory at least. This leads to believe that the algorithm used for the identification of communities hides part of the information. The latter is due to the fact that the algorithm is forced to make a choice between the two communities a node can be present in. If we allow for firms to be present in several communities we should observe some of the more central firms to be part of different communities. In figure 5.6 the network is represented at three points in time.

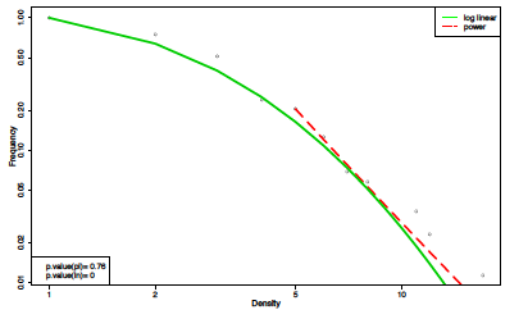
The giant component of the network (and even some of the smaller ones) clearly contains overlapping communities. As the network expands with time the number of communities increases. A cluster appeares around the year 2000 during the genomics boom (this can be seen more clearly in figure **??**). In the same timeframe the public research cluster appeares as well, including CNRS, INSERM and several universities. These two

[2]Source: http://www.total.com/sites/default/files/atoms/files/total-biomasse-en-final.pdf

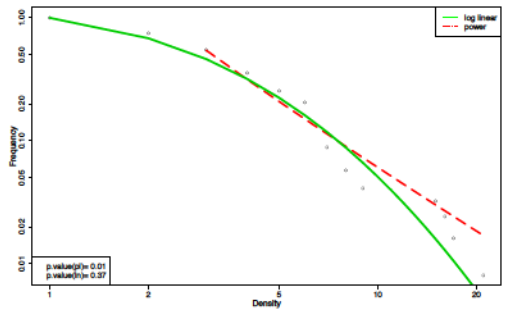


(a) Scale-Free and log-normal fit for 1988

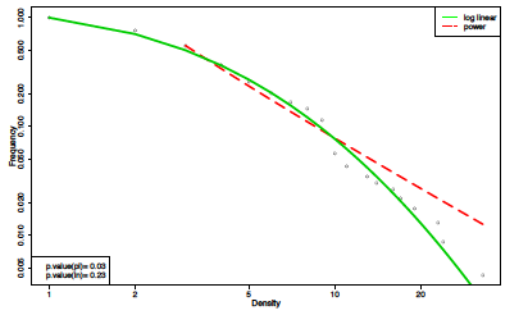
(b) Scale-Free and log-normal fit for 1992



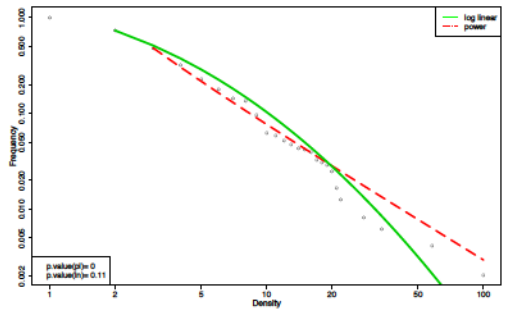
(c) Scale-Free and log-normal fit for 1996



(d) Scale-Free and log-normal fit for 2000

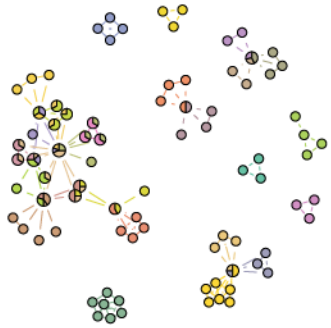


(e) Scale-Free and log-normal fit for 2004

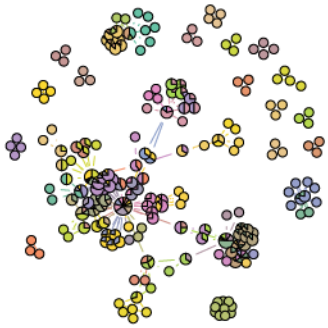


(f) Scale-Free and log-normal fit for 2013

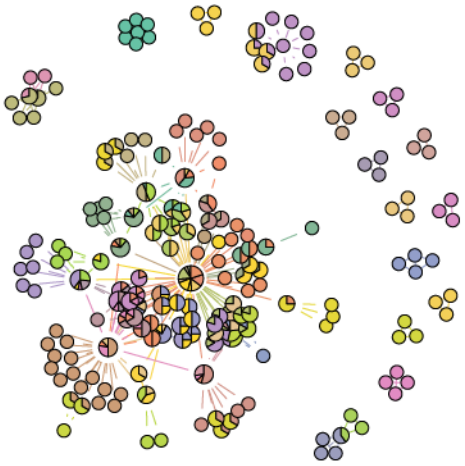
Figure 5.5: *Core-periphery fits for the Biotech sector in France between 1988 and 2013. The dotted line (red) represents the power-law fit while the full line (green) is the log-normal fit. The P.values of the fits are given in the lower left corner.*



(a) Communities 1980-1995



(b) Communities 1996-2005



(c) Communities 2006-2014

Figure 5.6: *Overlapping communities in the French Biotech sector in 1980-1995, 1996-2005, 2006-2014. The pie charts represent the percentages of membership to different communities. A pie chart that is cut in half implies that the firm is equally affiliated with both of the clusters it is in.*

clusters start to give shape to the network as a whole. When we look at the position of the more central firms in each cluster we can extract the overlapping communities. In figure 5.7 we can see that the CNRS is involved in many different, smaller communities. The cluster in which it is embedded is build up from a large variety of firms and universities which develops a large variety of technologies. The communities of Transgene and ARD are more precisely defined since their membership in one community exceeds 50%. These firms are more specialized than the other firms and hence evolve in their own well defined communities. The position of Biogemma appears to be less well defined as thought at first. It is present in its community for green biotechnology at 50% while also being present in the red biotechnology cluster of Sanofi-Avantis. The central firms in the different clusters interconnect the different types of biotech research in the sector. Smaller firms seem to be more specialized and are either present in the small components or in the smaller clusters at the periphery of the network. In order to verify this statement we will use the following method. Using the same patent database we will extract all the patents that were deposited by firms without collaboration (i.e one applicant on the patent). We extract the IPC codes of these patents and create an IPC-Applicant network. In other words a link is created between the applicant and each of the IPC codes on the patent. This will result in a bi-partite network in which clusters form either around a technology or around a firm. In the first case we have a fundamental technology for the sector that is mastered by many firms while the latter implies we have a firm that masters many technologies.

From this network we extract the minimum spanning tree in order to highlight the interconnection of these communities. We hence should be able to visualize when a firm is in many communities because it masters many technologies or if a firm is present in many communities because it has a specialized knowledge.

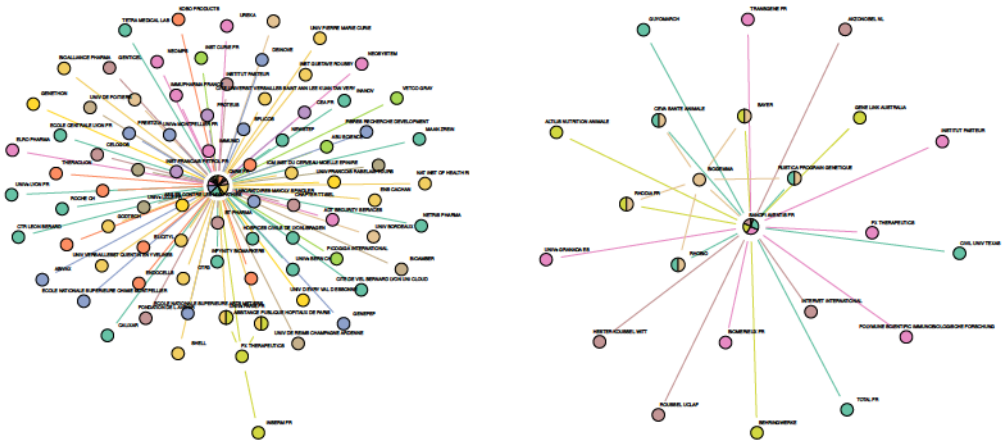
The larger firms collaborate to gain access to specialized knowledge from other firms in order to advance on their trajectory. This becomes visible when we look at the IPC-Firm network in figure 5.8. The most central nodes are either firms surrounded by a large diversity of technologies or technologies surrounded by the firms working on it. The diversity in technologies becomes clear for the CNRS which is directly connected to a relatively large number of different technologies. In addition it is a a low distance of other technology clusters. In contrast, Biogemma is connected to a low number of technologies, reflecting a higher level of specialization.

Figures 5.11(a) and 5.11(b) show that different clusters appear, some around technologies, some around firms. A61K for instance refers to biotechnologies and medicine while C12N are Micro-organisms or Enzymes. These codes are clearly central to the sector. We also observe several central firms surrounded by technologies (note that in this network only firms and technologies can be linked, links between firms or between technologies are non-existent). We can identify the firms we found in the community matrix within this network. In the complete network (Figure 5.11(b))The firms with the most technologies are the more central firms in the collaboration network.

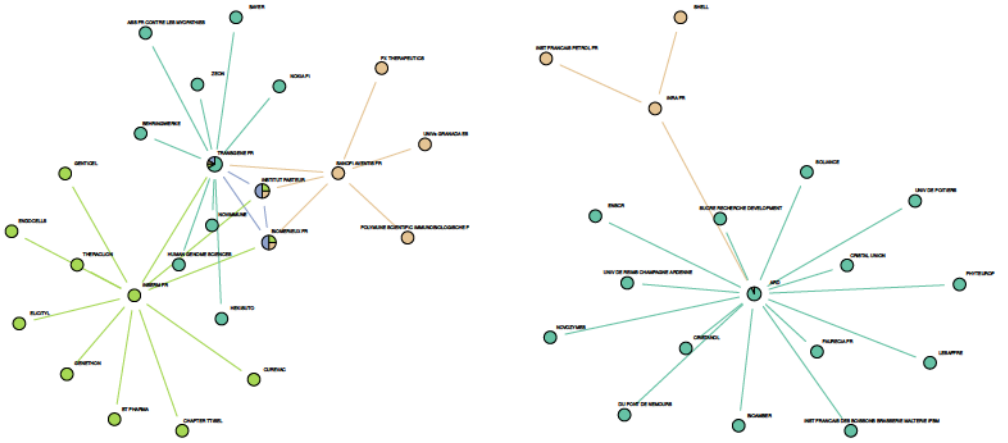
The Minimum Spanning Tree, as shown in figure 5.11(a), shows that several firms act as gatekeepers connecting different technologies, this is the case for Rhodia, Bayer and Innotherapie Lab for instance. These firms do however not have a central position nor are they present in different communities in the collaboration network. Rhodia makes the connection between two communities in the collaboration network, between Flamel technologies and Sanofi Aventis. There hence seems to be a correlation between the position in the technology network and the position in the collaboration network. As one would expect the larger firms to be able to master more technologies and to be able to sustain more collaborations explaining their central position in the network. This seems to be the case in this network. Small firms have more specific knowledge that create links between the larger firms. Indeed, the interest of research institutions for biotech from the year 2000 on changes the structure of the network by interconnection with many of the larger firms.

The presence of the Research Institutions (RIs) changes the landscape of collaboration since in the last period of the analysis the agents present in most communities were almost exclusively RIs. They have a central position in the network and collaborate with all different types of biotechnology clusters. When we look at their position in the knowledge network in figure 5.8, which connects firms with IPC codes, we can see the CNRS works on many different technologies. This results in a large cluster of public research that interconnects other clusters, densifying the network as a whole.

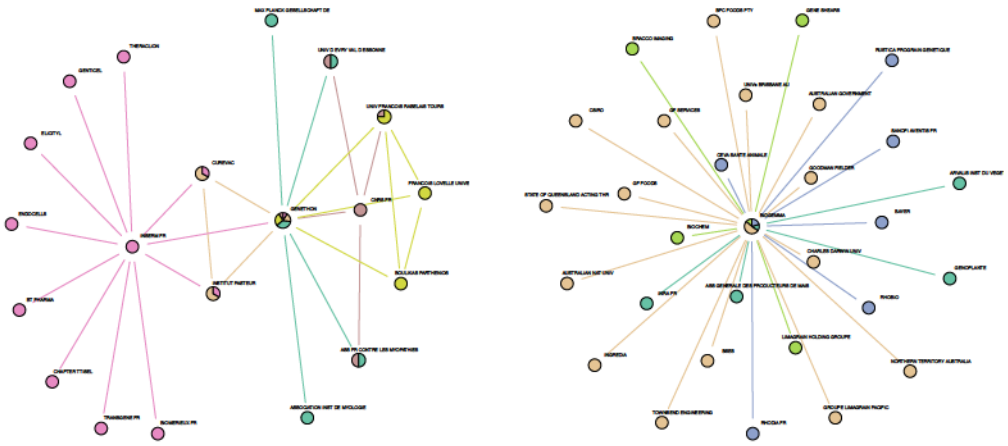
The very clearly defined cluster around Biogemma in the collaboration network is defined around different technologies mastered by the firms inside the cluster. In other words the cluster can be explained by the technologies mastered by the firms inside it.



(a) Overlapping communities of the CNRS (b) Overlapping communities of Sanofi-Aventis



(c) Overlapping communities of Transgene (d) Overlapping communities of ARD



(e) Overlapping communities of Genethon (f) Overlapping communities of Biogemma

Figure 5.7: *Overlapping communities of the firms with a central position in their communities. The pie charts represent the percentages of membership to different communities (identified by the colors of the pie chart).*

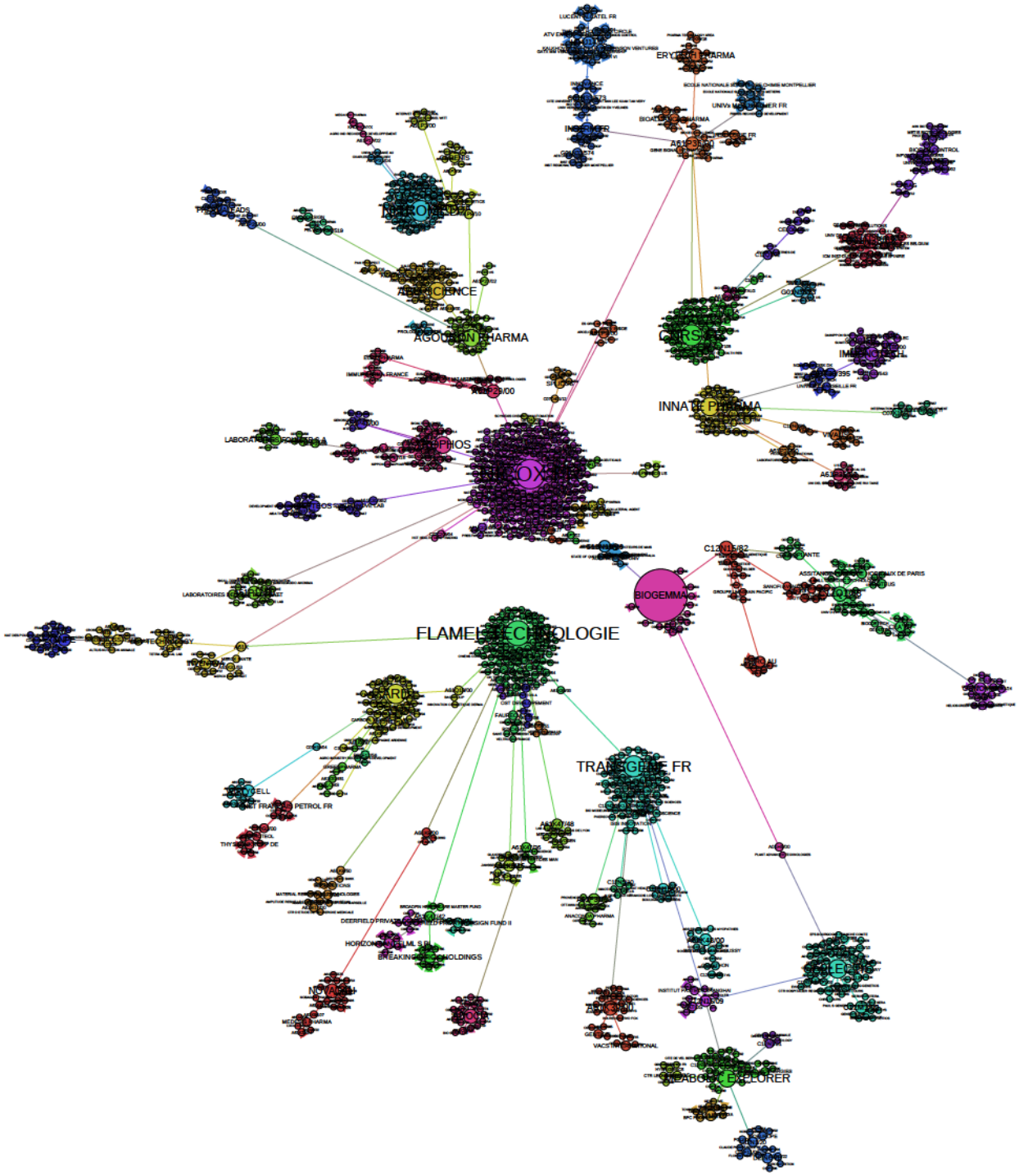
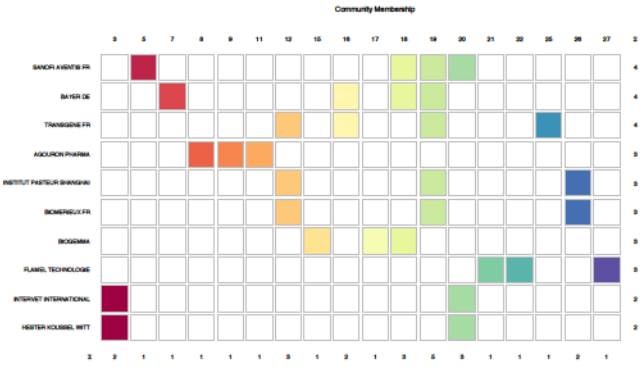


Figure 5.8: *Minimum Spanning Tree of the firm-IPC network of the french biotechnology network*

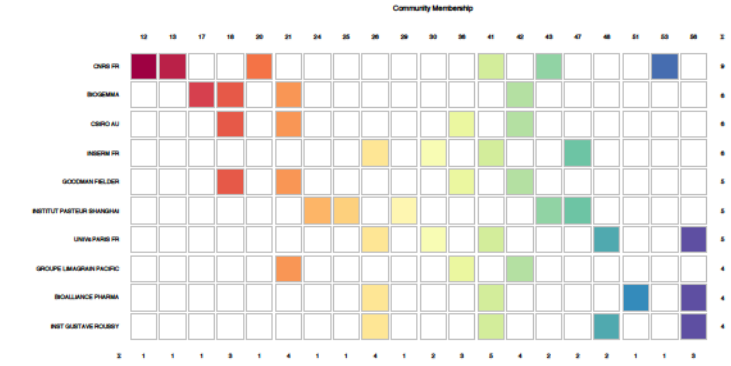
5.3.3 Community dynamics

The more central nodes are the ones that are present in most communities. These positions are occupied by either research institutions (NCRS, INSERM, universities) or large firms (Biogemma, Sanofi Aventis, Bayer). While in the first period (1980-1995) firms were the actors present in the most communities, the landscape completely changed in the second period with the introduction of Biogemma and the development of research institutions 5.9(a), 5.9(b). These Figures show the firms involved in the most communities in the network. The number on the right gives the number of communities, the number on the top gives the community ID while the number on the bottom gives the number of the identified firms are in the community.

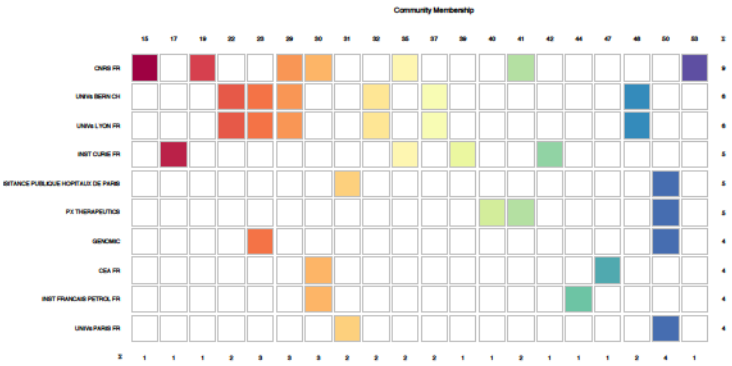
These matrixes show that firms were present in different communities while also have some of them in common (#12, #18 and # 19 for instance). This leads to believe that these firms work on different technologies.



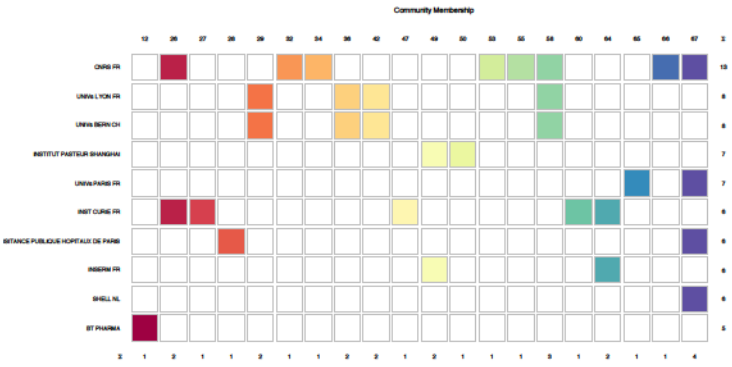
(a) Communities 1980-1995



(b) Communities 1996-2005



(c) Communities 2006-2014



(d) Communities 1980-2014

Figure 5.9: *Communities 1980-1995, 1996-2005, 2006-2014*

In order to understand to a broader extent the motivations of link creations between firms in the Biotech sector we use an ERGM model. For a detailed explanation of ERGM models please refer to chapter 2 of this thesis.

5.3.4 ERGM

An exponential Random Graph Model is used to identify the micro-level mechanisms explaining the overall structure of the network.

The results of the model can be found in table 5.1. A first observation is that both the edges and triangles are significant. Just as for the Aerospace sector, we can conclude that there is a significant impact of triadic closure (i.e a tendency to close open triangles). The intra-community collaborations tend to create triangles. There is hence a tendency for collaborators of collaborators to work together. The gwdegree variable is used to weigh the degree distribution to overcome degeneracy problems. The fact that it is significant shows that there is a need to compensate for the influence of the highly connected nodes in the network. This highlights the fact that the firms at the center of their communities play an important role in the structuring of the network. Finally, the proximity variable is also significant, as is the case in the aerospace sector. This leads to the conclusion that there is an inverted U-Shape relation between technological proximity and the tendency to collaborate.

5.4 Financial analysis for biotech sector

The results of the regression are in table 5.2.

We will compare the results of this regression with the results from the aerospace sector in order to highlight the impact of sectoral differences.

The adjusted clustering measure has a significant positive impact on the performance of the firm. These results would suggest that presence in clusters rather than connecting bridges is more efficient. The social capital theory seems to be validated, in other words the redundancy of information flow is outperformed by the impact of the increased efficiency resulting from working with firms that already know each-other.

The number of cooperations has an insignificant impact in the aerospace sector, it has a significant negative impact, (note that the number of cooperations is not equal to the

	Dependent variable:			
	Network			
	(1)	(2)	(3)	(4)
edges		−5.431***	−7.685***	−7.429***
		(0.124)	(0.298)	(0.006)
triangle			2.009***	1.736***
			(0.002)	(0.00001)
degree2	10.940***	0.426*	8.223***	1.528***
	(0.365)	(0.225)	(1.771)	(0.148)
degree3	11.820***	−0.006	3.176***	2.230***
	(0.462)	(0.215)	(1.072)	(0.134)
degree4	8.939***	−0.926***	8.122***	2.535***
	(0.781)	(0.226)	(1.790)	(0.226)
degree5	2.919***	−1.600***	4.038**	2.706***
	(0.942)	(0.258)	(1.568)	(0.322)
degree6	−0.332	−1.520***	6.770***	3.125***
	(0.778)	(0.262)	(1.411)	(0.311)
degree7	−4.771***	−2.521***	1.899	1.857***
	(1.533)	(0.406)	(1.455)	(0.611)
degree8	−1.030	−0.972***	4.026**	2.444***
	(0.916)	(0.221)	(1.754)	(0.318)
gwdegree			9.189***	
			(2.205)	
gwdegree.decay			−0.593***	
			(0.060)	
edgecov.proximity2	−2.629***	1.625***	1.959***	1.746***
	(0.006)	(0.132)	(0.350)	(0.008)
Akaike Inf. Crit.	11.909	8.370	7.337	7.161
Bayesian Inf. Crit.	11.981	8.451	7.445	7.251
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5.1: *Regression Results of the ERGM model.*

number of links).

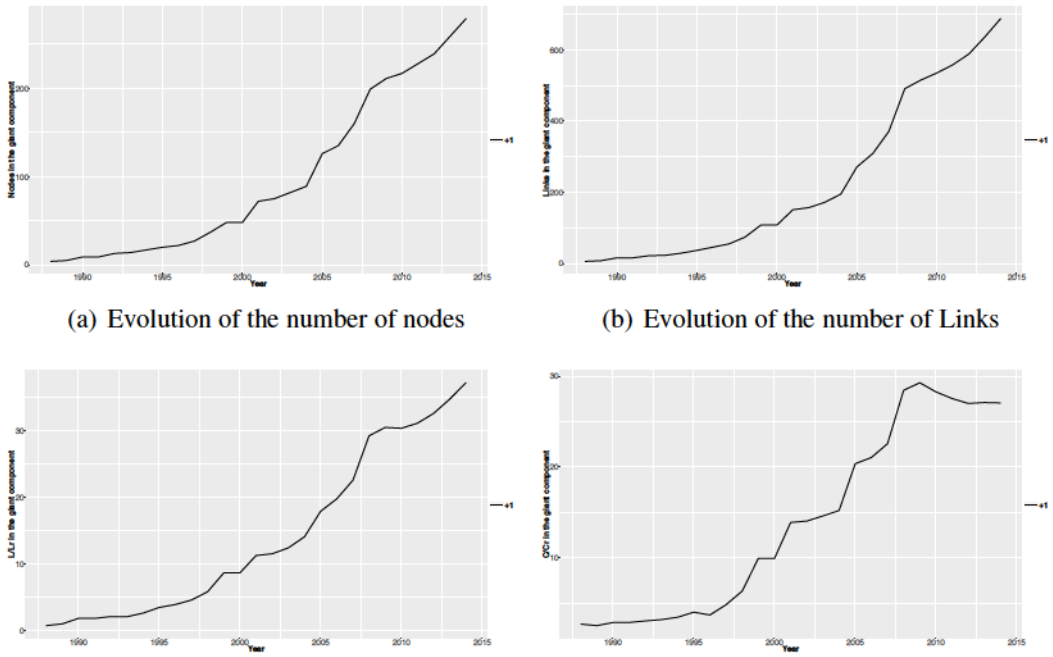
The number of technologies (absorption capacity) has a negative impact in both sectors even though the impact is higher in the biotech sector.

The most striking difference appears to be the insignificance of the average distance and the centrality. In the aerospace sector centrality had a positive impact on the performance of the firm while this is not the case in the biotech sector. The central firms in the aerospace network had a specific position in the value chain. In the biotech sector, the central firms are larger firms embedded in different technological clusters as shown by the network analysis. The fact that average distance and centrality do not have a significant impact on performance reflects the fact that firms are mostly impacted by their direct collaborators and competitors. The aerospace sector requires all clusters to communicate for the purpose of efficient knowledge transfer. This is not the case in the Biotech sector. Clusters develop their own specific technologies that do not require knowledge to flow between clusters. The interconnections of clusters is merely the result of firms that work on different types of biotechnologies. In other words, diversified firms. This diversity of technologies has a positive impact on performance. In addition, firms with a larger diversity of technologies in their neighborhood also benefit from an increased performance. The central firms in the network, which are the larger firms appear to be outperformed by the smaller more specialized firms.

	Model 1
Intercept	−3.39 (1.93)*
Clustering	−11.11 (3.94)**
Deg x Clust	2.24 (1.13)**
diversity	0.05 (0.01)***
Number of patents	0.24 (0.22)
Number of cooperation	−0.43 (0.26)*
Centrality	0.00 (0.00)
Average Distance	0.08 (0.85)
Number of technologies	−1.56 (0.40)***
factor(Year)2006	3.28 (2.64)
factor(Year)2007	3.39 (2.59)
factor(Year)2008	−0.40 (2.52)
factor(Year)2009	0.31 (2.46)
factor(Year)2010	−2.67 (2.42)
factor(Year)2011	−4.29 (2.39)*
factor(Year)2012	−4.92 (2.35)**
factor(Year)2013	−6.30 (2.37)**
factor(Year)2014	1.10 (4.70)
R^2	0.03
Adj. R^2	0.02
Num. obs.	3198
RMSE	29.36

$^{***}p < 0.001$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 5.2: *Panel regression results*



(a) Evolution of the number of nodes

(b) Evolution of the number of Links

(c) Evolution of the adjusted average distance coefficient

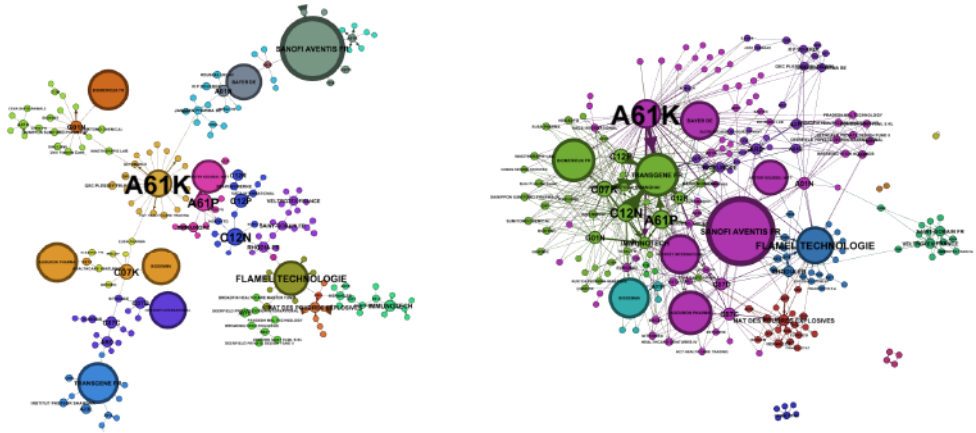
(d) Evolution of the adjusted clustering coefficient

Figure 5.10: *Structural dynamics of the French Biotech collaboration network between 1986 and 2013.*

5.5 Conclusion

The overall structure of the network is ill defined, no canonical structure could be identified. Hypothesis 1a. cannot be accepted. The structure of the network does not concur with any canonical structure. The structure is however highly correlated with the organization of the sector. The different types of biotechnologies are present in their own communities. The overlapping community identification has shown that the larger, diversified firms interconnect the different communities, giving the network the structure that it has. The ERGM model shows that technological proximity plays a similar role in both the biotech sector as the aerospace sector. There is an inverted U-shape relation between the technological proximity of the firms and the probability to collaborate, validating hypothesis 1b.

The ERGM model also shows that triadic closure plays a defining role in the structuring of the network, validating hypothesis 1c. The latter shows that collaborators of collaborators have a tendency to collaborate. This can be explained by the idea that firms collaborations in the biotech sector are risky due to the high level of competition. Firms open themselves



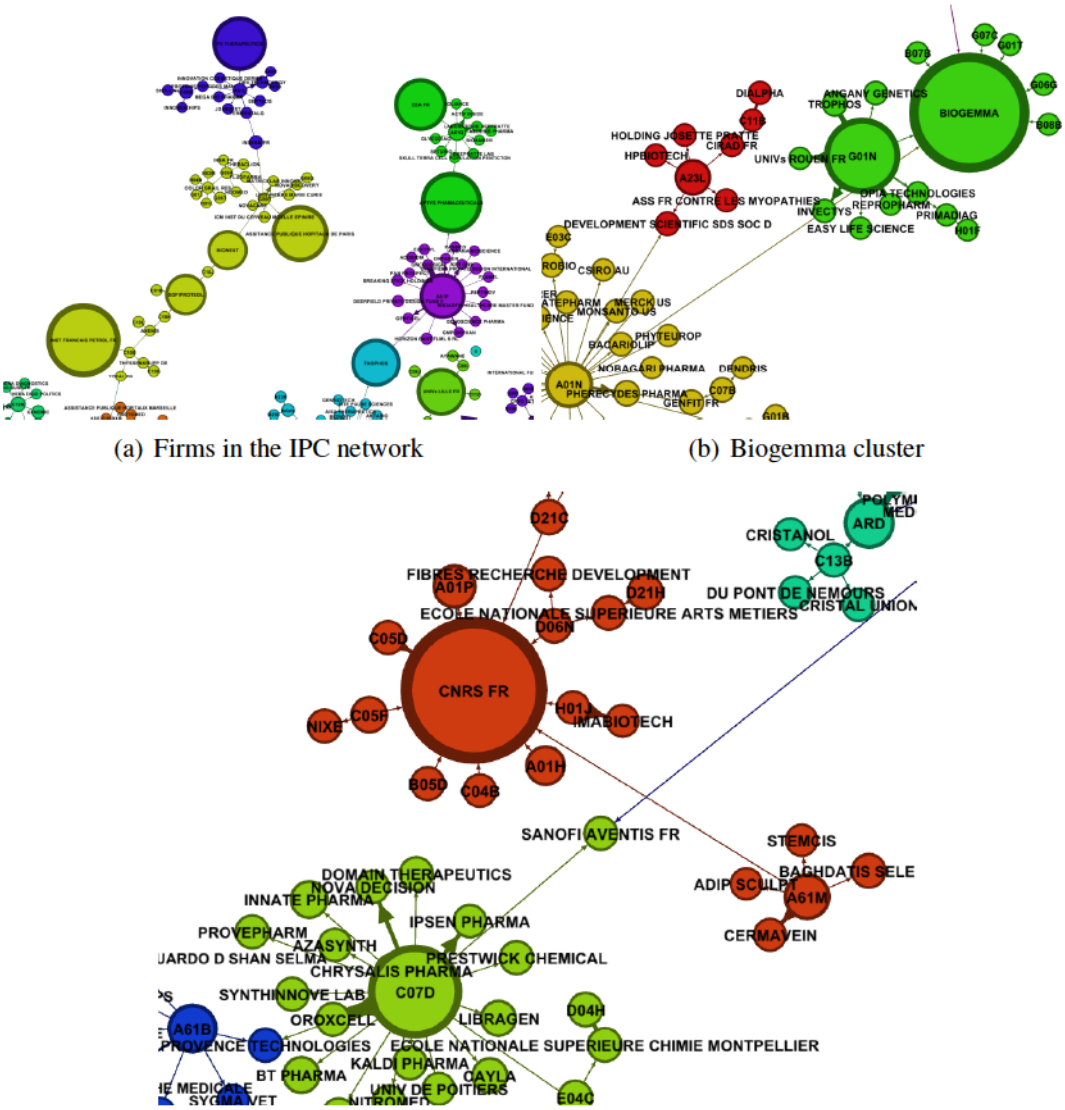
(a) Minimum spanning tree (b) Complete

Figure 5.11: *IPC-Firm networks for the Biotech sector 1980-1995*

up to the risk of opportunistic behavior of the other firms.

Overall there appears to be a correlation between the knowledge network of the sector and the collaboration network. The structure of the different clusters in the collaboration network can be explained by the knowledge mastered by the firms. Large firms with a large knowledge base have communities of firms with specific knowledge surrounding them. These observations are verified over 3 periods.

In the last period, the structure of the network is very much defined by the public research institutions with the CNRS interconnecting all the clusters. In this sector, defined by a high level of competition, collaborations are risky. Firms open themselves up to the risk of opportunistic behavior of the other firms. On the performance level, a notable difference between the biotech sector and the aerospace sector resides in the non-significance of some of the structural variables. Since there is little to no need for knowledge to flow through the whole network, the average distance between nodes and the rest of the network is not relevant. The most important factor is the neighborhood of the firm. In short, the position in the network does not matter as much as the quality of the neighborhood of the firm.



(a) Firms in the IPC network

(b) Biogemma cluster

(c) CNRS cluster

Figure 5.12: *Technology clusters in the last period*

Part III

Part C: Modeling Innovation networks

Chapter 6

Networks, knowledge dynamics and firm performance: Can firms have *too many* connections?

"There's something that doesn't make sense. Let's go and poke it with a stick." – The second Doctor

The previous chapters have answered questions relating to the link between the position of firms in the network and their performance. The question of the performance of the network itself has been left unanswered. It is difficult to assess the performance of a network empirically. Indeed, knowledge flows are difficult to measure, their impact even more. In order to overcome these obstacles, a theoretical model seems to be the best solution. The question we aim to answer here is the identification of a structure that outperformance other structures. In other words, is there a network structure that favors innovation?
As stated in the previous chapters, the value of networks resides in the flow of knowledge between firms. Internalizing the knowledge flows a firm is exposed to is of vital importance for the innovative abilities of the firm. One of the main reasons knowledge flows are studied is to understand how they impact the performance of either the firm or the network. In this light, it is vital to understand how the firm incorporated the knowledge it is exposed to into its R&D process. The environment of the firm has an important role to play here. The diversity of knowledge as well as technological proximity are important factors when it comes to knowledge diffusion and absorption. The model in this chapters aims at including

these aspects of knowledge diffusion. In addition, we explicitly model the manner in which firms use their external knowledge in order to increase their market shares, an aspect often missing from innovation network models.

By absorbing and sending knowledge to other collaborators, new technologies, ideas and information diffuses throughout the network (Leung, 2013). The structure of the network is then one of the foremost factors impacting the diffusion of knowledge (Verspagen and Duysters, 2004; Carayol et al., 2008; Cowan and Jonard, 2004). This has raised the question of optimal network structures from the point of view of knowledge diffusion. The theoretical models aiming at addressing this question have shown that dense networks with asymmetric degree distributions (Jackson and Wolinsky, 1996; Goyal and Joshi, 2003; Goyal and Moraga-Gonzalez, 2001; Cowan and Jonard, 2004) allow for optimal knowledge flows. In these models, optimality is studied by a comparison of the marginal (financial) cost of an additional link and the marginal benefit (utility, knowledge received) of the link. Results converge towards the observation that optimal networks are dense when the marginal cost is low. For a higher cost level intermediary structure emerge as optimal. König et al. (2012) extend this literature by showing that, for the same cost for links, different structures can be optimal. The structures they identified (the spanning star and a network comprised of several unconnected cliques of the same size) are however too regular to be considered empirically relevant.

It appears then, that the density of a network is one of the key features impacting knowledge flows. In this light we will focus our model on the impact of the density of collaboration networks on the diffusion of knowledge. To this end we will generate random network with an asymmetric degree distribution with a fixed average density.

The structure of the network, however important, is not the only factor affecting the diffusion of knowledge. Even if a firm is exposed to flows from a variety of sources, nothing guarantees that the firm is able to benefit from them. In order to properly internalize the knowledge into its innovation process the firms needs to be able to understand the knowledge. Savin and Egbetokun (2016) highlights the importance of the absorption capacity of firms in a network setting. The absorption capacity defines the extend to which firms

are able to internalize the knowledge they are exposed to (Cohen and Levinthal, 1990). If knowledge comes from firms that are too advanced, they will not understand each-other and no knowledge will flow. Taking this into account, firms can act as accelerators or obstacles slowing down the flow of knowledge. This follows the idea that knowledge decays as it flows through the network (König et al., 2009). So, in addition to the manner in which all firms are interconnected, the characteristics of different firms have to be taken into account as well. A model capable of including heterogeneity is hence a requirement.

This chapter is organized as follows: a first section will present an agent based model that aims at analyzing the impact of network structure on both the technological and economic performance of the firms. The second section will discuss the method used for the analysis of the model. The final section shows the impact of the structure on technological progress and economic performance. A final section is dedicated to the analysis of the impact of imitation.

6.1 The model

The aim of this model is to shed light on the impact of knowledge flows on technological and the economic performance of firms inside a network. We consider a model in which n firms produce and sell a homogenous good. We assume that firms use only capital for production. Our model consequently neglects, at this stage, the role of labor as a vector of knowledge transfer between firms.

We start from results already obtained in Jonard and Yildizoglu (1999) on the connection between network externalities, and extend this model in several directions, in order to better characterize the role of the network connections on the innovation process of firms, and on the dynamics of their industry.

Jonard and Yildizoglu (1999) is an industrial dynamics model inspired by Nelson and Winter (1982), and already contains some network interactions between innovation processes of firms. We extend this model by introducing a possibility for firms to discover new technological trajectories through radical innovations, and by opening the black-box of network externalities considered by the original model. Firms are now connected

in a horizontal innovation network. The links between firms allow them to exchange knowledge, and to reinforce their R&D processes with this knowledge. The impact of these knowledge flows on the R&D process of the firm depends on the technological diversity of their neighbors, and the technological distance between them.

Innovation allows firms to improve their production technology, and increase the productivity of their capital. With a higher productivity level, a firm can increase its output, and hence its market share. T These increased market shares increase the resources available for R&D investment and for innovation, consequently driving the growth of firms.

These activities take place, in each period of the model, following a sequence:

1. The current level of productivity of the firm results from past R&D, innovations and imitations
2. Firms produce, given their productivity and their capital stock
3. The production of the firm is sold at the current market price
4. Firms obtain their profits, and add them to their wealth
5. Firms decide on the amount they wish to invest in R&D, in physical capital and on the financial market
6. The investment in R&D allows the firms to innovate which can result in the discovery of new technologies or trajectories.

Our presentation of the model will mainly follow this sequence of operations.

6.1.1 Production and profits

Firms use physical capital as the only input for production. The supply of firm i at period t is given by the use of its capital $(K_{i,t})$ with the productivity $(A_{i,t})$ corresponding to its actual technological level:

$$y_{i,t} = A_{i,t} \cdot K_{i,t} \tag{6.1}$$

Total supply in the economy is given by:

$$Y_t = \sum_{i=1}^{n} A_{i,t} \cdot K_{i,t} \tag{6.2}$$

The quantity put on the market by firms faces a given demand represented by a constant elasticity, inverse demand function, and the current market price is fixed through the intra–period equilibrium on the market, as in Nelson and Winter (1982):

$$p_t = \frac{Demand}{Y_t^{\eta}} \tag{6.3}$$

where η represents the elasticity of inverse demand. Considering a fixed unit using cost of capital, c, the gross profit of the firm, at the end of the market process, is given by:

$$\pi_{i,t} = p_t \cdot y_{i,t} - c \cdot K_{i,t} \tag{6.4}$$

Each firm will dedicate these resources to investment in order to develop its activities for the next period.

6.1.2 Investments, technical progress, and transition to the next period

As in Jonard and Yildizoglu (1999), we assume that firms are perfectly aware that they are in a competitive environment where their survival strongly depends on their R&D activities. Consequently, investment in R&D is a priority for the firms. Investment in physical capital and on the financial market only takes place if enough resources are left after the R&D investment, which must consequently be considered in the first place.

R&D investment

Each firm invests a fraction (β) of its cash flow in R&D. This investment rate depends on the position of the firm in the market. The firm considers that it is lagging behind its industry if its market share becomes below the average market share ($1/n$). In this case, the firm will want to increase its R&D effort in order to catch up with the industry. We simple formulate such a behavior in the following manner.

Each firm starts with the same fraction, β_0, and increase its effort proportionally to its lag, in respect to the average market share:

$$\beta_{i,t} = \beta_0 + (1 + \zeta \cdot min(0, \frac{1}{n_t} - \lambda_{i,t})) \tag{6.5}$$

where $\lambda_{i,t}$ is the current market share of the firm, and ζ is the sensitivity of its R&D strategy to its lag with respect the average market share.

In addition to this variable investment we also consider that a fixed investment (r_{min}) is necessary to keep afloat the R&D laboratory. The total desired R&D investment of the firm is hence given by:

$$RD^*_{i,t} = \beta \cdot y_{i,t} \cdot p_t + r_{min} \tag{6.6}$$

We follow the idea that the knowledge held by the firm has a cumulative nature (Dosi, 1982) which increases with the R&D investment of the firm. But because some knowledge becomes obsolete in each period this stock can also deprecate, following a fixed obsolescence rate. The evolution of the knowledge stock of the firm is given by:

$$KS_{i,t} = (1 - depreciation) * KS_{t-1} + RD_{i,t} \tag{6.7}$$

This simple formulation is quite similar to the knowledge accumulation dynamics adopted in other articles (König et al., 2012; König, 2011).

The knowledge stock of the firm allows for the discovery of new technologies, through innovations.

Incremental and radical innovation

In this chapter we consider a long term evolution, and include consequently two types of innovations. Incremental innovation allows firms to discover new technologies along their current technological trajectory (Dosi, 1982), while radical innovation allows firms to discover a new trajectory. Figure 6.1 represents these two types of exploration of the technology space. ω represents the number of steps the firms can take when it discovers new technologies on its current trajectory, thanks to incremental innovations, and Δ represents the number steps the firm can take in the trajectory space, thanks to radical innovations.

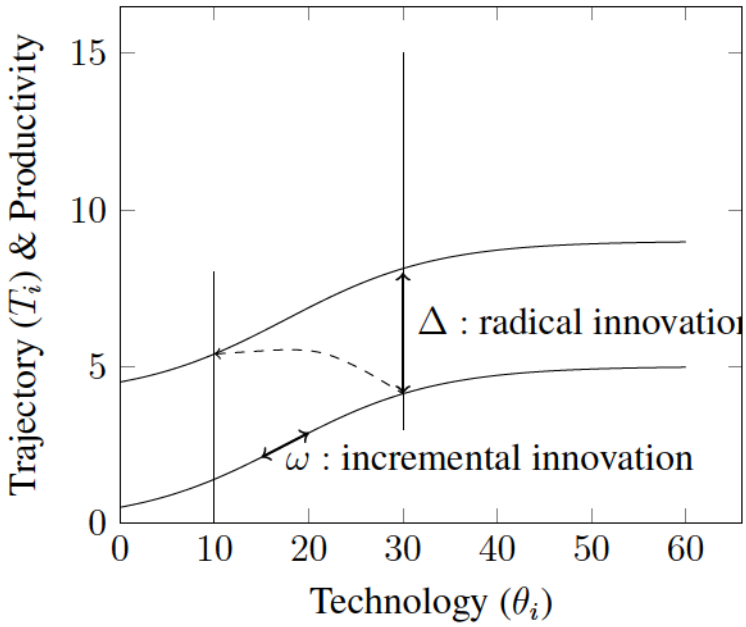


Figure 6.1: *Incremental and radical innovation mechanisms*

Each subsequent trajectory is a translation of the productivities corresponding to different technological levels (θ) with a given constant factor that corresponds to the size of radical innovations in this model.

Firms' discoveries depend on the knowledge stock into which they can tap. When firms are isolated, they can only rely to their own knowledge stock, while firms connected to other firms can benefit from the knowledge stock of their neighbors. But, this *knowledge effect* of network connections does not play the same role for both types of innovations. Not all knowledge that flows in its network is relevant for the firm, it depends on the type of innovation it is searching for.

Incremental innovation For the discovery of incremental innovations, the firm can only benefit from knowledge that is relevant to the current trajectory that it is exploring. The firm can use its own knowledge and the knowledge of neighbors with the same (or a better) technology that evolve on the same trajectory. But firms must be able to absorb the knowledge. The more advanced the neighbors of the firm, the more difficult for it to understand their technological knowledge. Consequently, the network effect in this case (NE_{inc}), must aggregate the knowledge of the neighbors taking into account this absorption capacity that decreases with the distance between technologies. At the limit, the firm can benefit fully from the technology of its neighbor, under the condition that they

both use the same technology. We also consider that a minimal amount of autonomous knowledge flow is always present (and represented by γ_{inc}):

$$NE_{inc,i,t} = \frac{1}{\#\mathcal{N}_{i,T_i=T_j}} \cdot \sum_{j \in \mathcal{N}_i, T_j = T_i} \left[\left(\gamma_{inc} + 1 - e^{max-1/max(0,\theta_j - \theta_i)} \right) \cdot KS_{j,t-1} \right] \quad (6.8)$$

where $\#\mathcal{N}_{i,T_i=T_j}$ is the number of neighbors who use the same trajectory as the firm.

Based on these elements, we can define the knowledge pool (KP) into which the firm can tap for its incremental innovations:

$$KP_{i,t} = \begin{cases} KS_{i,t}, & \text{if } \#\mathcal{N}_{i,T_i=T_j} = 0 \\ KS_{i,t} + NE_{inc,i,t}, & \text{if } \#\mathcal{N}_{i,T_i=T_j} > 0 \end{cases} \quad (6.9)$$

This knowledge pool helps the firm to take new steps in the technology space, over its current technology. We assume that the number of steps it may take follows a Poisson distribution with a mean equal to $log\left(KP_{inc,i,t}\right)$. When the drawn value is 0, no discovery is made. When the draw is strictly positive, it gives the number of steps $(\omega_{i,t})$ the firm will take, from a starting technology that depends on the nature of the knowledge process in its industry: if the innovations are collectively realized, the starting point will depend on the average technology of the neighbors of the firm that are on its trajectory $\left(\overline{\theta}_i\right)$. We represent this dependence on the social dimension of the incremental innovation process using a parameter α_{inc} and, consequently, the new technology that will be discovered by the firm will be given by:

$$\theta_{i,t} = \left[(1 - \alpha_{inc}) \cdot \theta_{i,t-1} + \alpha_{inc} \cdot \bar{\theta}_i \right] + \omega_{i,t} \quad (6.10)$$

Consequently, when the firm has only neighbors that are lagging behind it, the social dimension of the incremental innovations can impede her ability to discover new technologies. We will call this effect of the network connections the *lock–in effect.*

There is a point on the trajectory at which the firm starts to face serious decreasing returns to innovation, i.e the innovations start providing an increasingly smaller increase in productivity. When a firm observes that its R&D investments are not providing important enough improvements on the current trajectory, it can decide to search for a better techno-

logical trajectory, i.e. to innovate radically. The idea is that a firm will decide to switch from incremental innovation to radical innovation once the end of its current trajectory has been reached.

Radical innovation When the firm is not able to consistently increase its productivity through incremental innovation, even after several (winProd) periods, it starts to explore new trajectories and dedicate all its resources to radical RD. This switch happens once the firm notices that it has reached the end of its current trajectory. During this exploration process it will continue to exploit the last technology discovered on its current trajectory. Given the size of the inventive step necessary for discovering radical innovations, we assume that searching for radical innovation and new trajectories necessitates a stronger combinatory process and larger knowledge base, coming from much broader sources than in the case of incremental innovation.

Hence, technological diversity plays an essential role in this case (Sampson, 2007; Miller, 2006; Suzuki and Kodama, 2004). Consequently, knowledge flows from firms on different trajectories, as well as from firms with different technologies on the same trajectory contribute to the knowledge pool of the firm for radical innovation, and we consider that the farther is a neighbor, in terms of trajectory or technology, the stronger its contribution to the knowledge base of the firm. The weight of trajectory–diversity v.s. technology–diversity on the same trajectory is represented by a parameter, ν , the higher the value of the parameter the more importance is given to the technology–diversity. In order to measure the contribution of each neighbor, j, to the knowledge basis of the firm i, we build a distance indicator combining both sources of diversity:

$$dist_{i,j} = \sqrt{\upsilon \cdot (t_j - t_i)^2 + (1 - \upsilon) \cdot (T_j - T_i)^2}$$

Where t_i is the technology of firm i and T_i the trajectory of firm i. Using this distance index, we can now compute the network effects from which the firm benefits in its radical innovations.

$$NE_{rad,i,t} = \frac{1}{\#\mathcal{N}_i} \sum_{j \in \mathcal{N}_i^d} (\gamma_{rad} + (1 - e^{\frac{-10}{dist_{i,j}}})) \cdot KS_{j,t-1}$$

where $\#\mathcal{N}_i$ is again the number of neighbors of the firm and γ_rad is a minimal network

effect that is always present. If $dist_ij = 0$, firm j does not contribute to the knowledge base of i since its knowledge is redundant with i's knowledge. Based on these elements, we can define the knowledge pool (KP) into which the firm can tap for its radical innovations:

$$KP_{rad,i,t} = \begin{cases} KS_{i,t}, & \text{if } \#\mathcal{N}_i = 0 \\ KS_{i,t} + NE_{rad,i,t}, & \text{if } \#\mathcal{N}_i > 0 \end{cases}$$

This knowledge pool helps now the firm to make new vertical steps in the trajectory space (radical innovations). As before, we assume that the number of steps it may do follows a Poisson distribution with a mean equal to $log\left(KP_{rad,i,t}\right)$. When the realized value is 0, the firm does not discover a new trajectory at all. When the draw is strictly positive, it gives the number of steps (Δ)the firm will make, from a starting trajectory that depends on the nature of the knowledge process in its industry: if the innovations are collectively realized, the starting point will depend on the average trajectory of the neighbors of the firm $\left(\bar{T}_{\mathcal{N}i}\right)$. We represent this dependence on the social dimension of the radical innovation process using the parameter α_{rad} and, consequently, the new trajectory that will be discovered by the firm will be given by:

$$T_{i,t} = \left[\alpha_{rad} \cdot T_{i,t-1} + (1 - \alpha_{rad}) \cdot \bar{T}_{\mathcal{N}i}\right] + \Delta$$

A firm may however not adopt the new trajectory right away, depending on its initial productivity. Indeed, as shown in figure 6.1, the firm starts at the beginning of a new trajectory (since it does not yet have any experience with it), and the corresponding productivity level might be lower than the current one, with the old trajectory. The firm start then to exploit its new trajectory, by investing in incremental RD for this new trajectory, and will switch to it only when it has reached a productivity level at least equal to the current one, on the old trajectory. It will then adopt the new trajectory and start to use the last technology discovered on it.

Imitation

From time to time (with a probability $probImitate$), firms are able to observe how their neighbors operate, as well as their trajectories and technologies. When the imitation is successful, the firm copies the trajectory and technology of the neighbor with the highest

productivity, and is able to attain immediately the same productivity as the copied firm. The imitation process as we introduce it in this chapter is quite fortunate, and optimistic for the lagging firms. We prefer to keep it very simple, and just as a possibility, at this stage, because we plan to introduce a finer modeling of appropriability conditions and of a patent system in a subsequent project. Our main results in this version will be developed below in the version of our model without any imitation, and we will introduce this possibility just as a variant to check if imitation can bring supplementary structure to the distribution of firms in the technology space.

6.1.3 Capital investment and exit conditions

We use the same procedure for capital investment as inJonard and Yildizoglu (1999), and the exit condition is based on the resources available to the firm.

Physical capital investment

The firm acknowledges that RD is a condition of survival in this very competitive Nelson Winter industry, and uses its profit and savings in order to finance its RD activities as a priority. But the amount the firm desires to invest in RD can exceed the current profit and savings of the firm, it may have to sell part of its physical capital on a scrap market to reach RD^*. The sale of capital will be completed at a scrap rate. Formally, if $\pi_{i,t} + Savings_{i,t} < RD^*_{i,t}$ the firm will have to sell an amount equal to $RD^* - (\pi_{i,t} + Savings_{i,t})$ at a scrapping price s. The quantity of capital sold is given by:

$$s \cdot \Delta K = max\left\{0, RD^* - (\pi_{i,t} + Savings_{i,t})\right\}$$

If the firm still has profit and saving left after investing in RD it will either invest on the financial market where it can earn an interest rate r or invest in capital to increase its productive ability. If the return capital (ρ) exceeds the interest rate it will invest in capital and vice-verse. The firm invests in physical capital when:

$$\rho - \delta \geq r$$

Where δ is the depreciation of capital, and ρ is computed based on the expected profit of the next period using a linear regression (see Jonard and Yildizoglu (1999)). In addition, it is supposed that firms observe decreasing returns to investment. If we note I_t the investment of the period then the firms estimate their profit for the next period by:

$$\hat{\pi}_{i,t+1} = \hat{\alpha}_i \cdot \sqrt{K_{i,t}} + \hat{b}_i$$

This gives the investment criterion:

$$\rho = \frac{\hat{\alpha}_i \cdot \sqrt{K_{i,t}} + \hat{b}_i}{K_{i,t} + I_t} - \delta \geq r$$

The firm invests in capital as long as the return exceeds the interest rate. The investment is then $I_{i,t}$. Finally, the new level of capital once the investment is completed is given by:

$$K_{i,t+1} = (1 - \delta)K_{i,t} + I_{i,t}$$

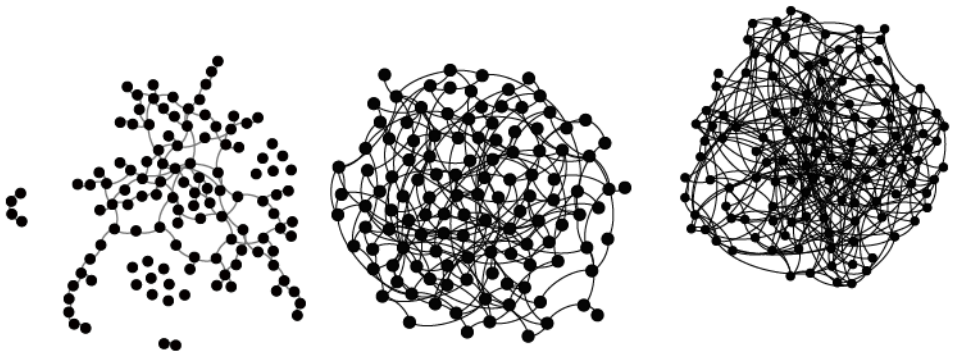
Exit

Firms exit the industry when they are left without even a minimal level of physical capital ($K_i <= survivalCapital$).

6.2 Simulation protocol

We adopt an exploration of the parameter space based on the Nearly Orthogonal Latin Hypercubes (NOLH) design of experiments (DoE) (Salle and Yıldızoğlu, 2014) for analyzing the results given by the model on our main research question: the role of network density in the technical progress, and in the structuring of the industry. The NOLH panel we create for our main parameters contains 65 experiment points, and we repeat each of them 40 times, for 300 periods. That gives us 2600 runs of 300 periods in total for each network configuration. We analyze results coming from the distributions of indicators at T=300. We stop at period 300, because for later periods, exits appear and change considerably the structure of industry between configurations, make them difficult to compare. Even without exits, we can compare different concentration levels between network configurations (see below). The networks/industries that we study contain 121

firms, and each of them starts the history on trajectory 0 and on technology 0, and with the corresponding productivity. Their physical capital stock is initialized at initial-capital. Their initial innovation type is incremental. The values of parameters explored in our experiments are given in Appendix E.



(a) Generated network with an average density of 2 (b) Generated network with an average density of 4 (c) Generated network with an average density of 6

Figure 6.2: *Generated networks for different values for the average distance.*

6.2.1 Generating networks with different densities

We keep the network structure exogenous in this model, and we fix it at the beginning of the simulation. Lets note g a network. The presence (or absence) of a link between two nodes i,j in the network g is defined by a binary variable g_{ij} taking value 0 if a link is absent and value 1 if a link is present. We then define the degree of a node i as the number of nodes to which it is connected:

$$d_i = \sum_{j=1}^{n} g_{ij} \tag{6.11}$$

The average degree of the network is then the average of the degrees of the individual firms:

$$\bar{d}_g = \frac{1}{n} \sum_{i=1}^{n} \sum_{i \neq j}^{n} g_{ij} \tag{6.12}$$

Since our focus is on the impact of network density on industrial and technological performance, we generate and compare networks with different fixed values for $\bar{d}_g$. To ensure having an identical level of average density we use the following simple algorithm:

1. Pick one node at random and create a link with another random node;
2. Compute the average density of the graph;
3. If the average density is higher or equal to the desired density, stop. Else, repeat from step 1.

Once the desired density has been reached the algorithm stops. This results are random networks with a given average density. Some examples of these generated networks can be found in figure 6.2. For each level of density considered in this model (0,1,2,3,4,6 and 8) we ran the simulation 2600 times. A density 0 is used as a benchmark case without any network effects. We compare the technological and industrial performance between these network configurations using different indicators.

6.2.2 Indicators and measures

Besides standard indicators like average and maximal productivity, and economics variables like the market price, of profits generated by the model, we introduce two sets of more dedicated indicators. One concerns the structure of the networks, and the second the measure of the industrial concentration.

Centrality of firms in the network

Centrality is a measure of the position of a node inside a network. More precisely, it aims to measure how well connected a node is inside a network. Since economic theory provides different uses for network analysis, there are different measures of centrality, each providing an answer to a different question. In the case of this chapter we aim at identifying the importance of a firm according to its access to knowledge from the rest of the network. The centrality measure needs to include a measure of the different paths a node is positioned on. The more paths the node is on the more access to knowledge it has (directly and indirectly, in time). As such we chose to use Betweenness centrality as a measure of centrality. Betweenness centrality measures the paths a node is positioned on, and hence provides a measure of how well it is positioned in terms of information flows in the network. For a given node it checks the number of shortest paths the firm i is positioned on:

$$\sum_{k \neq j,\ i \in \{k,j\}} \frac{\frac{P_i(kj)}{P(kj)}}{\frac{(n-1)(n-2)}{2}}$$

$P(kj)$ is the number of shortest paths between nodes k and j, and $P_i(kj)$ the number of shortest paths on which the focal node, i, is positioned. All the possible paths are summed and averaged over the total number of possible paths $\left(\frac{(n-1)(n-2)}{2}\right)$. The shorter the distance from one node to all the other nodes in the network, the higher the centrality.

Measuring concentration: Inverse–Herfindahl Indexes

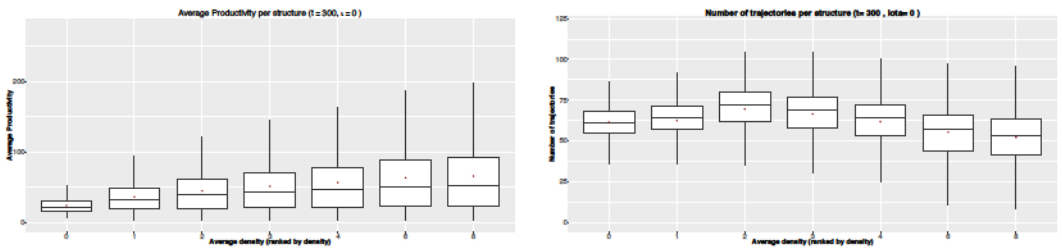
In order to get an idea of the distribution of market shares of firms, we use an inverse–Herfindahl index. This values ranges from one (all demand is addressed by one firm) to the number of firms (the market is equally shared between all firms). The index is computed as follows:

$$H_Q = \frac{(\sum_{i=1}^{n} y_{i,t})^2}{\sum_{i=1}^{n} y_{i,t}^2} \tag{6.13}$$

We also compute a similar indicator for the concentration of the physical capital between firms:

$$H_K = \frac{(\sum_{i=1}^{n} K_{i,t})^2}{\sum_{i=1}^{n} K_{i,t}^2} \tag{6.14}$$

6.3 Results

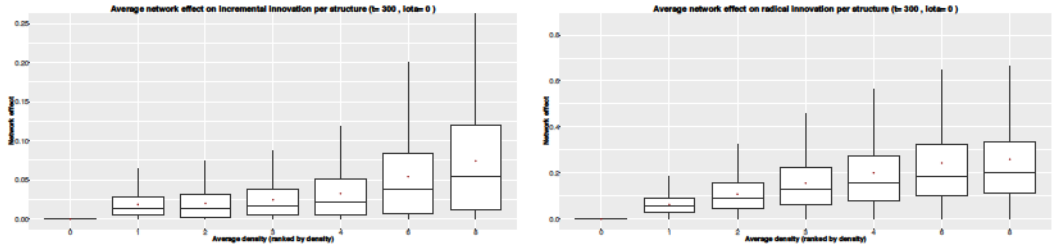


(a) Average productivity of firms for $\iota = 0$ (b) Average number of trajectories with $\iota = 0$

Figure 6.3: *Average productivity and the average number of trajectories per network density, without imitation ($\iota = 0$)*

Using this model that incorporates multiple potential roles that networks can play in technology and industry dynamics, we analyze the importance of the density of these networks from the point of view of the firms, and of social welfare. We first focus our attention on technology dynamics, and how they can be influenced by the density of the network. We then put in perspective technological dynamics by connected them with firms' performance, industrial dynamics, and social welfare. This analysis is conducted under

pure network effects, without taking into account the imitation of technologies between firms in the network. The last subsection is dedicated to the role played by imitation.



(a) Average network effects on incremental innovation with $\iota = 0$

(b) Average network effects on radical innovation with $\iota = 0$

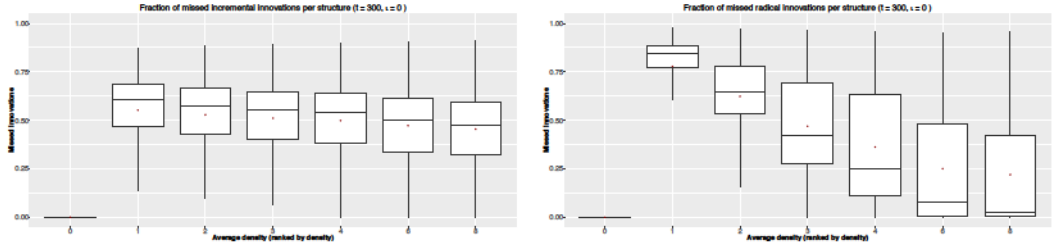
Figure 6.4: *Network density, and contribution of network effects on incremental and radical innovation propensity of firms, without imitation* ($\iota = 0$)

6.3.1 Network density and technical progress

We compare the distributions of average productivities corresponding to different network structures at $t = 300$ to assess their consequences on the average technical progress in the industry. The $0-$density case is used as a benchmark industry without any network effects (empty network), in which firms each evolve alone. Figure 6.3(a) shows that the presence of networks are favorable to technical progress, and higher densities drive a stronger collective technical progress. But we also observe that this positive effect gets smaller when the density increases. In fact, using the non–parametric Wilcoxon-Mann-Whithney (**WMW**) tests, we can even determine that this effect becomes negligible for $d \geq 3$ (we cannot reject the null hypotheses of homogeneity of distributions between all cases corresponding to $d \geq 3$, for $\alpha = 5\%$)[1]. Consequently, if the connections are costly, and they are indeed in many industrial contexts, the firms would prefer to belong to networks with a positive, but relatively low level of density. Depending on this cost, we can expect, from the technological perspective, an optimal network density corresponding to a degree $d = 1$ or $d = 2$.

What explains the decreasing returns from network density ? Even if the contribution of the network effects on the knowledge stock of the firms (the *knowledge effect*), and on their ability to discover new technologies/trajectories through innovation, remain increasing

[1] Using WMW tests, we can order the medians of average productivity distributions (observed in $t = 300$) in networks corresponding to different densities, d: $8 = 6 = 4 = 3 > 2 > 1 > 0$.

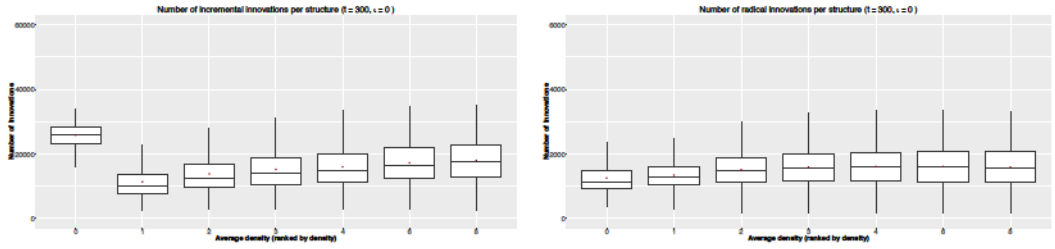


(a) Rate of missed incremental innovations per structure with $\iota = 0$

(b) Rate of missed radical innovations per structure with $\iota = 0$

Figure 6.5: *Rate of missed innovation without imitation* ($\iota = 0$)

with network density (Figures 6.4), the final technological levels of the firms are also subject to a stronger *lock-in effect* that keeps firms on a lower number of trajectories (Figure 6.3(b)). As in Jonard and Yildizoglu (1999), a stronger network effect is a source of lower technological diversity, even in the presence of radical innovations. This *lock–in effect* clearly appears in Figures 6.5(a)-6.5(b) that show the proportion of discovered innovations that have not been adopted because they were not superior to the technology used by the innovating firm. This possibility arises when the discoveries of the firm depend on the average technological level of their neighbors (measured by the parameter $socialDimension$).



(a) Number of incremental innovations

(b) Number of radical innovations

Figure 6.6: *Number of innovations without imitation* ($\iota = 0$)

This lock–in effect is the strongest when the networks of firms are the most localized ($d = 1$). It decreases when the size of their networks increases, and their network includes a higher share of the industry, but the lock-in effect remains significant even for $d = 1$. The decrease is steeper for the radical innovations where the discovery of new technologies depends on the average trajectory of the neighbors: since firms are concentrated on smaller numbers of trajectories when the average density is high (Figure 6.3(b)), the dependence on average trajectory is not really crippling for their radical innovations.

Another way of looking at this contrast between the network effects on incremental and

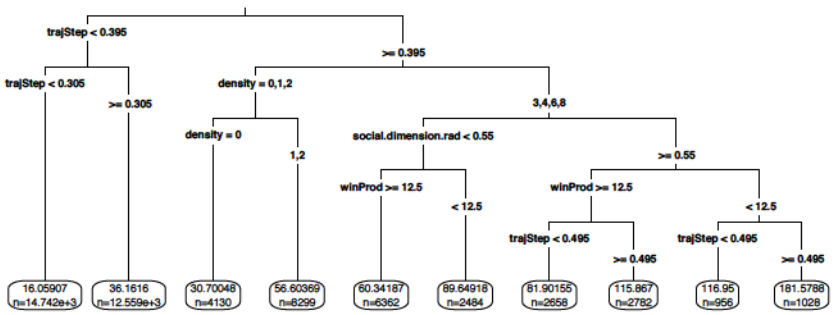


Figure 6.7: *Regression tree for the average productivity* ($cp = 1\%$)

radical innovations is in terms of successful innovations (Figures 6.6): Network effects reduce the number of incremental innovations, in comparison with the empty network case, even if this negative effect decreases with their density, while their presence clearly supports a higher number of radical innovations.

We can hence conclude that the final positive effect on average productivity must result from this positive effect on radical innovations, and dimensions of technical progress that are particularly favorable to that type of innovations should play a significant role in the determination of the average productivity of firms.

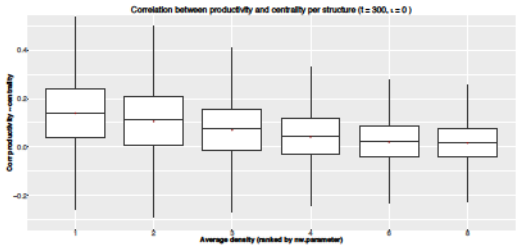
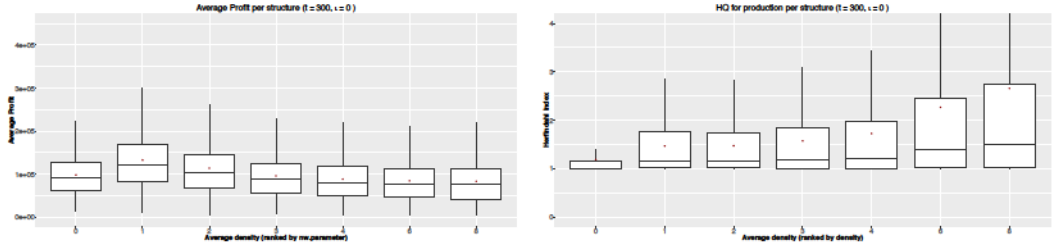


Figure 6.8: *Correlations of productivity of firms with their centrality (without imitation,* $t = 300$*)*

In order to check this result we develop a regression tree analysis Venables and Ripley (2013) where the dependent variable is the average of the average productivity. A regression tree partitions the set of observations using the relative contribution of different explanatory variables on the expected value of the dependent variable. The final leaves of the tree give the expected value of the dependent variable in the corresponding subset of observations, and the number of the observations. We see in Figure 6.7 that the most prominent variable for average productivity is the size of radical innovations ($trajStep$): we observe the



(a) Distribution of average profits (b) Distribution of inverse Herfindahl indices

Figure 6.9: *Network density, profits and concentration without imitation* ($\iota = 0$)

lowest expected average productivity (16.059) when it is low (< 0.395), and other factors are not able to compensate this weakness of the radical innovation process. At the other end of the distribution, we observe the highest expected average productivity (181.58) when the size of radical innovations is higher (≥ 0.495), the network density is not too low ($d \geq 3$), the social dimension of radical innovations is not too low (≥ 0.55) ,and the memory of the firms, in assessing their performance in incremental innovations, before deciding to switch to the radical ones, is not too high ($winProd < 12.5$). When the density of the network is too low ($d < 3$), other dimensions of the innovation process cannot compensate this weakness, and the expected average productivity is at most equal to 56.60.

At the individual level, we can also observe, in Figure 6.8, that firms who play the role of the *connector* for the others in the network (higher centrality) benefit from higher technical progress: the productivity of firms is mainly positively correlated with their centrality, especially in networks where the connections are less dense. This effect is the strongest when the connections are rare ($d = 1$), and when the density increases, and all firms end by having numerous neighbors in the network, this effect levels off, and the correlations become closer to zero.

Consequently the density of firms' network, and the dimensions of the process of radical innovations play in a complementary way and support the technological performance of the firms. Are these factors also favorable to the economic performance of firms?

6.3.2 Network density and economic performance

Technological performance is globally higher in denser networks. Does a higher network density necessarily result in a higher economic performance for the firms and the social welfare? If not, what would be the socially optimal density level, even without

taking into account the potential cost of connections? In order to answer these questions, we consider the distributions of the average profit of the firms, and of the market price (which is inversely related to the consumers' surplus, because the demand is given in our case).

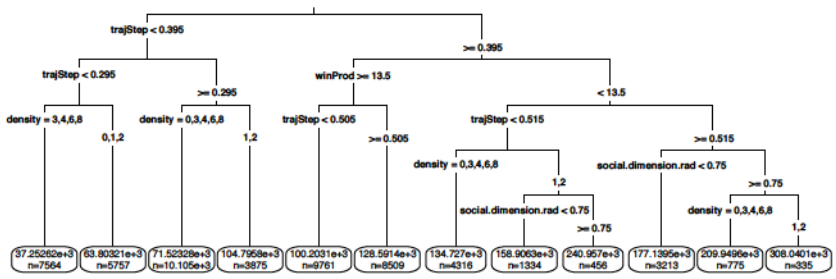


Figure 6.10: *Regression tree for the average profits* ($cp = 1\%$)

We clearly observe in Figure 6.9(a) a non monotonic relation between the average profits and the density of networks: network connections help firms to obtain higher profits, but only when they are not *too* dense. Ordering the medians of these distributions for different degrees of density, d, using WMW tests give the following ordering: $1 > 2 > 0 > 3 > 4 > 6 > 8$. This result indicates that it is better for firms to not have any network than having too many connections. So, form the point of view of the economic performance as well, firms would prefer less dense networks here. This is true even without taking into account the cost of connections. We could think that firms selection would be weaker in a more densely connected network because all firms benefit, one way or another, from the knowledge stocks of the others. This effect is indeed confirmed by the Figure 6.9(b) that shows that the concentration clearly decreases with the density of the networks. But this relation is rather monotonous and cannot explain the evolution of the profits. We can note that this evolution is quite in-line with the ability of the firms to explore different trajectories (Figure 6.3(b)), and this diversity is a major source of knowledge for the quality of radical innovations.

In order to verify this potential explanation, and better understand the determinants of the average profits, we use again a regression tree where the dependent variable is the average profits (Figure 6.10). The most significant factor is again the size of radical

innovations ($trajStep$), but in all configurations, $d = 1, 2$ dominate other densities in terms of average profits. We can confirm here that the highest profits are indeed obtained through the interaction of the low densities with high social dimension (≥ 0.75) in radical innovations: with low but positive densities, firms are able to benefit from the knowledge effects, while suffering from limited lock-in effects in radical innovations. Hence, in these cases the articulation of these effects is more favorable for the firms. All branches resulting in these highest profits point to the role of radical innovations ($trajStep \geq 0.515, winProd < 14, social.dimension.rad \geq 0.75$).

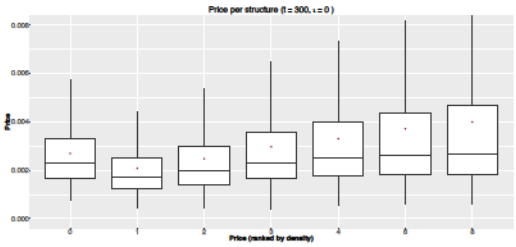
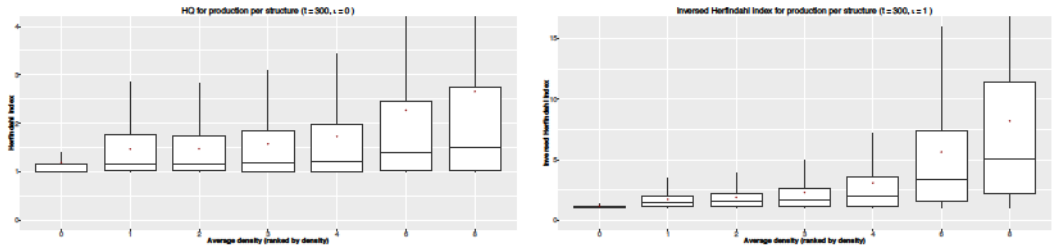


Figure 6.11: *Distributions of market price (without imitation, $t = 300$)*

Are these higher profits obtained at the disadvantage of the consumers? Concentrations increasing with the density of networks would let to think us that lower densities would be unfavorable to consumers' welfare. But, we can directly check the effect on consumers' surplus here, because, for a given demand, the lower the prices, the higher the surplus. Figure 6.11 shows that this surplus is maximal for low but positive densities ($d = 1, 2$). So, such configurations seem to clearly drive a stronger technical progress, better profits, and higher consumers' surplus and, consequently, a higher social welfare. The relationship between network density and the economic performance is definitely not monotonically increasing, because of the complex articulation between the knowledge, selection, and lock-in effects of network connections of firms.

6.3.3 The effects of imitation

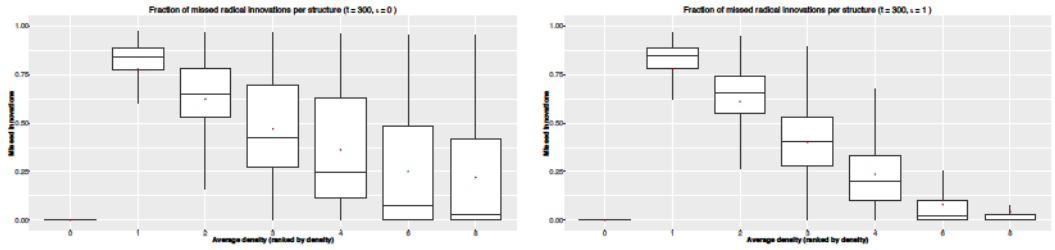
Until now, we have only considered the structuring of knowledge dynamics through innovations. But connections with other firms, especially competitors like here, may also serve the firms to better copy the technology of these competitors. In order to take into account the diffusion of technologies over the network, we now include the imitation process in our results ($\iota = 1$). Indeed, localized imitation could change the similarity



(a) Without imitation (b) With imitation

Figure 6.12: *Distribution of inverse Herfindahl indices without and with imitation*

between the firms and their neighbors, and considerably transform knowledge dynamics over networks. Also, by allowing a catching up by the lagging firms, it could increase their probability of surviving, and hence, the selection effect in the model. Does this possibility radically change our results?



(a) Without imitation (b) With imitation

Figure 6.13: *Rate of missed radical innovations without and with imitation*

Concerning the selection effect, Figures 6.12 clearly show that this effect is lower with imitation, and yield a lower concentration for the industry: without imitation, considering the firms that obtain the significant market shares, we have at best a duopoly in many simulations without imitation, while we can easily have more than 5 firms with imitation.

Concerning the technology dynamics, the most radical effect of imitation is observed for the missed radical innovations (Figures 6.13): by increasing the homogeneity of neighborhoods, imitation considerably reduces the lock-in effect in radical innovations in dense networks, and we have, globally, a much lower rate of unsuccessful radical innovations, when this density if high. We can consequently expect that the advantage of low densities over high densities should be much lower with imitation.

In order to check this conjecture, we now build a regression tree where the dependent variable is the average productivity, and we include now also the possibility of imitation ($\iota = 1$) in the explanatory factor (Figure 6.14). First we observe that radical

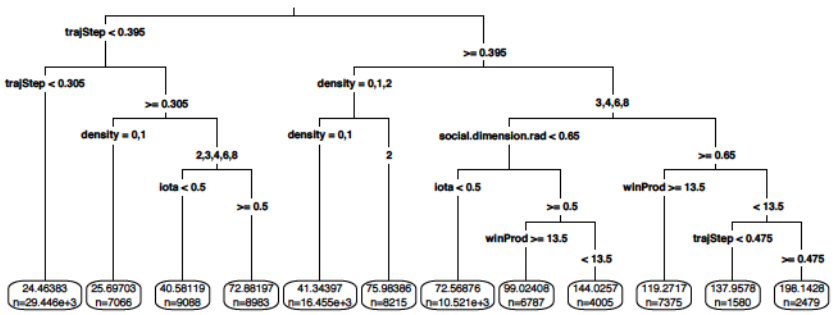


Figure 6.14: *Regression tree for the average productivity, including imitation* ($cp = 1\%$)

innovations (their characteristics) continue to be the most important driver of technical progress. In compliance with our conjecture, imitation only becomes discriminant when networks are dense ($d \geq 3$): only in these cases $\iota = 1$ may yield a higher expected average productivity in some configurations, when $trajStep \in [0.305, 0.395[$ or $trajStep \geq 0.395, social.dimension.rad < 0.65$. When we are already in configurations favorable to radical innovations, the effect of diffusion becomes less discriminant.

6.4 Conclusion

This chapter has focused on the impact of network density on the performance of the network. The model shows that the presence of networks is favorable for innovation as measured by the level of technological progress. However, firms observe decreasing returns to network density. Therefore, networks with low levels of density appear to be optimal. Too many links create a lock-in effect that hinders innovation. This effects increases when the number of links per firm increases. In the case of dense networks, the influence of the network increases and it becomes more complicated for firms to find new technologies because they are kept back by their collaborators. This is especially true for industries in which technological progress advances at a fast pace. When networks are scarcely connected and technological progress is swift, the highest performance is observed. From an economic point of view we have shown that there is a non-monotonic relation between the average profit in the network and network density. When density increases, the diffusion of knowledge results in firm selection becoming less aggressive.

The model shows that optimal performance is achieved for industries with fast innovation, in low density networks in which there is an important social dimension. This combination of conditions also provides the highest consumer surplus. Imitation allows firms to catch up easily and hence lowers industry concentration. In addition imitation allows for the reduction of the lock-in effect that dampening the economic and technological performance of the firms.

Part IV

Conclusion

Conclusion

This thesis is based on three main questions: how can we explain and interpret the structure of a collaboration network? do firms with specific positions inside the network outperform those with a less favorable position? And finally, are there network structures that are more favorable for innovation? My first conclusion is that I now have more questions than when I started.

The first question, is answered with three empirical studies. We have learnt that, at the level of the technology, the structural dynamics of the collaboration network are highly correlated to those of the life-cycle of the technology. An interesting observation is that towards the end of the period, firms start leaving the collaboration network of SCM since they start to focus on a new technology. The structure of the collaboration network at the level of the sector stabilizes over time. So, if firms stop working on a specific technology they still appear to be working with the same firms on other projects. If this were not the case the dynamics of the collaboration network would have been much more hectic. We can deduce from this that even if the technology evolves, the firms keep collaborating with mainly the same firms. This idea is supported by the clusters in the network. In the aerospace sector, each cluster represents a part of the aircraft while in the biotech sector each cluster represents a tier of the market. These clusters were well defined and quite stable over time. Innovation is localized in these networks. The different factors that rule partner choice go into this direction as well. The probability of cooperation between two firms follows an inverted U-shape with the level of technological proximity. Firms inside the same cluster are hence technologically close. Triadic closure is also significant, implying that firms with a common collaborator have a higher probability of collaboration than firms without a common collaborator. This observation coincides with the stabilization of the different clusters in the network. The same strategies appear to work inside the clusters in both the aerospace and the biotech sector. The global network structure is however defined by the

manner in which the clusters are interconnected and that is ruled by sectoral specificities. In the case of the biotech sector the CNRS interconnected the different clusters directly. In the aerospace sector Airbus connected with its suppliers which are connected to their clusters. The conclusion around the life-cycle was only tested on one technology. In order to make the results more robust testing the theory on different technologies is required. For the sectoral analyses I would like to expand the ERGM analysis by including more firm level data to identify more link creation strategies.

The second question, do firms with specific positions inside the network outperform those with a less favorable position, was analyzed in two different sectors. The results highlight that there is a correlation between network position and performance of the firm. This result is dependent on the sector of the analysis. In the case of the aerospace sector, the central position of the firm as well as its distance to the other firms have a positive impact on the performance of the firm. It would appear that firms which are exposed to more knowledge flows outperform those exposed to less knowledge flows. This effect is not significant in the case of the biotech sector. The competitive nature of the sector would keep knowledge from flowing outside of the clusters. And this clustering has a positive influence on performance in both sectors. We can recall here that triadic closure was identified as one of the strategies of the firms which seems to pay off. The specific environment of the firm, measured by the diversity of technologies in the neighborhood of the firm, has an important role to play for the performance of the firm. Firms with a large diversity of technologies in their direct neighborhood have an increased performance compared to firms with a lower diversity. That being said, this analysis needs work. Identifying the impact of knowledge flows is a delicate matter since the signal we try to identify is weak. More data should improve the results. The problem with this analysis is that missing observations create a double biais. The usual selection biais appears in addition to a change in the network. Removing firms from the network can drastically alter the structure, resulting in false observations.

The final question, are there network structures that are more favorable for innovation, is answered using an agent-based model. The model shows that networks have a positive impact on performance. This impact decreases with the density of the network. The most

efficient networks, from the point of view of innovation and economic performance, are hence sparsely connected networks. This is particularly true in the case of industries with fast innovation and where the social dimension is important.
Both the networks that were studied empirically had low levels of average densities. In particular the aerospace network, which has been optimized by the power8 program, matches the criteria. The model neglects certain important aspects. It would be interesting to include a partner selection mechanism in order to endogenize the network. This would allow us to study how different technological regimes result in the emergence of different network structure. In addition, employee mobility has been neglected as a factor of knowledge transfer.

R is used to treat the data and compute the different indicators and networks in this thesis. I have compiled these scripts into an R-package. The package contains however much more than the indicators in this thesis. Through different projects around this thesis I have programmed many indicators for Science and Technology (proximity indicators, specific IPC networks, patent and publication citation indicators, patent thickets for example). A beta-version of the package (SciTechR) is available here.

Appendices

Appendix A

Normalizing constant computation

Computation of the normalizing constant of the power-law.

$\int_{-\infty}^{\infty} p(x) = 1 \Leftrightarrow \int_{+\infty}^{x_{min}} p(x)dx = 1$

$\Leftrightarrow c\int_{+\infty}^{x_{min}} x^{-\alpha}dx = 1$

$\Leftrightarrow \int_{+\infty}^{x_{min}} x^{-\alpha}dx = \frac{1}{c}$

$\Leftrightarrow \left[\frac{x^{1-\alpha}}{1-\alpha}\right]_{\infty}^{x_{min}} = \frac{1}{c}$

$\Leftrightarrow \frac{1}{(1-\alpha)}\left[x^{1-\alpha}\right]_{\infty}^{x_{min}} = \frac{1}{c}$

$\Leftrightarrow c(\alpha, x_{min}) = \frac{\alpha-1}{x_{min}^{\alpha-1}}$

Appendix B

Network indicators

In the different chapters of this thesis different network indicators were used both in the analytical and econometric analysis. This appendix provides the formulae and interpretation of the indicators.

Centrality Centrality is a measure of the position of a node inside a network. More precisely it aims to measure how *central* a node is inside a network. Since economic theory provides different uses for network analysis there are different measure of centrality, each providing an answer to a different question. In the case of this thesis we aim at identifying the importance of a firm according to its access to diverse knowledge. The centrality measure needs to include a measure of the different paths a node is positioned on. The more paths the node is on the more access to knowledge it has. As such we chose to use *Betweenness centrality* as a measure of centrality. Betweenness centrality measures the paths a node is positioned on and hence provides a measure of how well it is positioned in terms of informational flow. For a focal node it checks the number of shortest paths the firm is positioned on.

$$\sum_{k \neq j,:\ i \in \{k,j\}} \frac{\frac{P_i(kj)}{P(kj)}}{\frac{(n-1)(n-2)}{2}} \tag{B.1}$$

$P(kj)$ is the number of shortest paths between nodes k and j, and $P_i(kj)$ the number of shortest paths the focal node, i, is positioned on. All the possible paths are summed and averaged out over the number of possible paths from which there are $\frac{(n-1)(n-2)}{2}$. The shorter the distance from one node to all the other nodes in the network the higher the

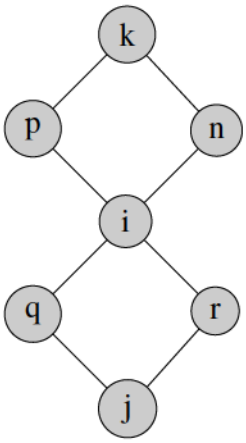


Figure B.1: *Brokerage and Closure illustration*

centrality.

Let's say we want to compute the centrality of node i. We then need to find the number of shortest paths between nodes all nodes and find on how many of these paths i is present. We will give an example of this computation for nodes j and k.
There are 4 shortest paths between nodes j and k:

1. j-q-i-p-k
2. j-q-i-n-k
3. j-r-i-n-k
4. j-r-i-p-k

$P(kj) = 4$ here since all the paths we found are the shortest possible in the graph. Node i is present on all those paths, hence $p_i(kj) = 4$. Node i is hence an important node for flows between nodes j and k, indeed, without node i the two would not even be connected. It is possible that a node is important for the interconnection of only a few nodes in a huge network. In order to weigh the importance of each possible link we normalize by computing the number of possible links in the network: $\frac{(n-1)(n-2)}{2} = \frac{6*5}{2} = 15$. We then find :

$$\frac{\frac{P_i(kj)}{P(kj)}}{\frac{(n-1)(n-2)}{2}} = \frac{\frac{4}{4}}{15} = \frac{1}{15} \tag{B.2}$$

This value is then computed for all possible links in the network giving a value for the importance of a node in a network. This measure is perfect for measuring the importance of a node when it comes to access to knowledge since it measure the shortest paths the nodes is on, and the shorter the path the more efficient the flow of knowledge, the more paths a firm is on, the higher the potential diversity of knowledge.

Clustering The clustering coefficient is a measure of cohesiveness in a network, in other words, how well connected the network is. The measure is quite simple; it represents the number of triangles in the network divided by the number of possible triangles.

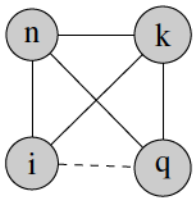


Figure B.2: *Clustering illustration*

Consider figure B.2, to find the clustering of the graph we need the number of triangles in the network. There are two triangles in the network: i-n-k and n-q-k. The number of triangles is hence equal to two. The number of possible triangles is equal to the number of triangles if the network was a complete network. The dotted link between nodes i and q makes the network a complete network. If this link existed we would have two additional triangles: i-n-q and i-k-q. The number of possible triangles is hence equal to four. The clustering coefficient is then equal to:

$$Clustering = \frac{\sum_{i,j\neq i,k\neq j,k\neq i} g_{ij} \cdot g_{ik} \cdot g_{jk}}{\sum_{i,j\neq i,k\neq j,k\neq i} g_{ij} \cdot g_{ik}} = \frac{2}{4} = 0.5 \tag{B.3}$$

The same value can be computed at the node level. This would give a measure of the extend to which firms' neighbors are connected. It gives the fraction of the neighbors that are connected.

Whether measure at the level of the node or the network level, the clustering coefficient gives a measure of embeddedness. When clustering equals one all possibles triangles exist, the more it tends towards zero the less triangles are observed.

Firms evolving in a highly clustered environment are at risk of a reduction of diversity

of knowledge since they cooperate with the same firms. Methods and ideas diffuse between the same firms and hence are less prone to new developments as would be the case for firms with less clustered neighborhoods. The latter would gain access to more new methods and ideas.

The clustering coefficient is hence a very important network statistic. It allows for the identification of clusters (densely connected areas) inside a network. At the node level it can identify if some nodes are shielded from the rest of the network (in terms of knowledge flows for example). A positive side effect of a highly clustered neighborhood is the common practices idea. Since all firms know each-other they work more efficiently together since they know about their work ethics and methods.

Degree The degree of a node is simply the number of distinct nodes a node is connected to.

Appendix C

P.values for the powerlaw fit

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0,57	0,92	0,064	0,018	25	3	3,207	2,152	0,982
1981	0,43	0,53	0,073	0,024	28	1	3,120	1,930	1,083
1982	0,8	0,79	0,059	0,020	29	1	3,359	1,930	1,115
1983	0,49	0,25	0,073	0,027	33	1	3,473	2,069	1,116
1984	0,63	0,6	0,060	0,021	32	1	3,087	2,103	1,095
1985	0,45	0,15	0,074	0,035	38	8	3,289	2,625	0,896
1986	0,77	0,56	0,054	0,027	36	7	3,409	2,710	0,889
1987	0,14	0,62	0,081	0,028	36	11	3,385	2,992	0,784
1988	0,01	0,96	0,075	0,022	21	9	2,686	2,846	0,839
1989	0,01	0	0,086	0,038	23	1	2,800	2,350	1,104
1990	0	0,53	0,091	0,021	26	1	2,866	2,273	1,089
1991	0,07	0,13	0,089	0,028	34	1	3,394	2,280	1,056
1992	0	0,25	0,101	0,026	14	1	2,341	2,287	1,063
1993	0,59	0,13	0,065	0,029	36	1	3,567	2,189	1,075
1994	0,42	0,76	0,067	0,018	37	1	3,672	2,267	1,070
1995	0,53	0,33	0,071	0,023	45	1	3,999	2,349	1,065
1996	0	0,38	0,081	0,022	12	3	2,280	2,520	0,982
1997	0	0,2	0,084	0,027	11	3	2,197	2,496	1,014
1998	0	0,6	0,088	0,020	14	3	2,243	2,572	1,029
1999	0,08	0,31	0,087	0,024	54	1	3,751	2,540	1,085
2000	0	0,11	0,076	0,028	15	2	2,307	2,567	1,055
2001	0	0,13	0,076	0,026	11	1	2,128	2,537	1,079
2002	0,56	0,41	0,066	0,022	57	2	3,911	2,594	1,056
2003	0,12	0,4	0,054	0,024	19	5	2,454	2,712	0,977
2004	0,06	0	0,056	0,035	20	3	2,484	2,724	1,003
2005	0,54	0,48	0,064	0,024	62	5	3,560	2,829	0,949
2006	0,07	0,17	0,057	0,028	21	3	2,512	2,790	0,977
2007	0	0,72	0,074	0,020	21	3	2,468	2,710	1,007
2008	0,09	0,75	0,061	0,022	20	9	2,477	2,477	1,068
2009	0,92	0,29	0,052	0,032	47	12	3,263	2,345	1,081
2010	0,41	0,69	0,067	0,024	35	8	3,153	2,164	1,065
2011	0,51	0,94	0,060	0,016	22	2	2,859	2,072	1,032
2012	0	0	0,027	0,009	41	24	2,292	0,805	1,819
2013	0	0,04	0,028	0,009	42	26	2,298	0,618	1,858
2014	0	0,05	0,028	0,009	42	26	2,298	0,621	1,858

Table C.1: *P.values for the 4-digit data (window)*

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0,03	0,01	0,074	0,044	13	1	2,875	1,821	1,055
1981	0,01	0	0,063	0,030	13	1	2,884	1,898	0,990
1982	0,6	0	0,031	0,030	15	1	2,777	1,920	1,060
1983	0,34	0,06	0,038	0,020	18	1	2,741	1,971	1,069
1984	0,52	0,01	0,032	0,020	20	1	2,694	2,041	1,055
1985	0,64	0,06	0,030	0,019	23	4	2,686	1,763	1,209
1986	0,11	0,12	0,040	0,017	24	5	2,591	1,489	1,333
1987	0,09	0,04	0,035	0,019	21	5	2,527	1,724	1,318
1988	0,2	0,51	0,039	0,011	33	5	2,648	1,762	1,318
1989	0,58	0,44	0,030	0,011	35	4	2,724	2,019	1,240
1990	0,13	0,11	0,040	0,013	39	5	2,713	2,037	1,265
1991	0,29	0,77	0,037	0,009	44	4	2,760	2,133	1,247
1992	0,02	0,77	0,045	0,008	42	4	2,636	2,139	1,259
1993	0,01	0,55	0,045	0,011	43	11	2,616	1,793	1,372
1994	0	0,54	0,047	0,011	27	13	2,454	1,562	1,435
1995	0	0,21	0,044	0,013	26	12	2,403	1,677	1,413
1996	0	0,14	0,038	0,014	28	13	2,437	1,687	1,424
1997	0,26	0,16	0,044	0,013	92	13	3,033	1,712	1,422
1998	0	0,29	0,040	0,012	27	13	2,390	1,791	1,406
1999	0	0,11	0,041	0,014	28	14	2,387	1,665	1,451
2000	0	0,35	0,039	0,011	27	13	2,359	1,923	1,388
2001	0	0,11	0,046	0,013	34	11	2,419	2,164	1,323
2002	0	0,12	0,039	0,012	31	11	2,381	2,224	1,326
2003	0	0,05	0,042	0,011	33	5	2,394	2,484	1,245
2004	0	0,04	0,038	0,012	34	6	2,383	2,461	1,268
2005	0	0,03	0,041	0,012	36	5	2,383	2,523	1,258
2006	0	0,05	0,037	0,010	34	4	2,358	2,598	1,242
2007	0	0,32	0,038	0,010	37	14	2,359	2,123	1,423
2008	0	0,4	0,040	0,010	39	14	2,370	2,276	1,390
2009	0	0,31	0,043	0,010	40	14	2,356	2,308	1,395
2010	0	0,23	0,038	0,011	38	17	2,337	2,196	1,435
2011	0	0,2	0,036	0,011	41	15	2,361	2,293	1,414
2012	0	0,07	0,035	0,010	39	4	2,334	2,704	1,276
2013	0	0,27	0,036	0,010	39	14	2,335	2,374	1,396
2014	0	0,22	0,036	0,010	39	14	2,334	2,374	1,396

Table C.2: *P.values for the 7-digit data (window)*

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0,09	0,08	0,031	0,014	28	8	2,764	1,589	1,249
1981	0,05	0,04	0,039	0,014	41	9	2,923	1,529	1,265
1982	0,02	0,01	0,036	0,016	30	9	2,800	1,430	1,292
1983	0	0	0,041	0,015	28	8	2,720	1,640	1,285
1984	0,74	0	0,029	0,020	61	10	3,184	1,233	1,397
1985	0,96	0	0,026	0,015	69	9	3,450	1,712	1,264
1986	0	0,01	0,036	0,016	35	12	2,902	2,273	1,117
1987	0	0	0,038	0,015	35	5	2,889	2,377	1,091
1988	0,02	0	0,037	0,019	44	5	3,043	2,398	1,045
1989	0,07	0	0,034	0,016	45	7	3,056	2,381	1,048
1990	0,26	0,2	0,039	0,024	72	50	3,305	3,001	0,970
1991	0	0,11	0,030	0,021	20	27	2,721	2,245	1,125
1992	0,04	0,55	0,034	0,015	38	28	2,898	2,068	1,153
1993	0,12	0,03	0,036	0,016	51	15	3,070	1,107	1,367
1994	0,18	0,05	0,034	0,015	51	15	3,089	0,855	1,411
1995	0,14	0,06	0,033	0,013	52	16	3,161	1,925	1,170
1996	0,02	0	0,031	0,013	43	7	2,985	2,387	1,034
1997	0,02	0	0,024	0,017	26	7	2,783	2,468	1,003
1998	0	0,36	0,027	0,015	24	31	2,669	1,439	1,338
1999	0	0,04	0,027	0,016	36	26	2,751	-0,388	1,694
2000	0	0,42	0,028	0,010	30	21	2,619	-0,141	1,699
2001	0	0	0,028	0,015	28	16	2,550	0,931	1,515
2002	0,01	0,14	0,026	0,010	33	17	2,553	0,306	1,672
2003	0	0,33	0,025	0,009	26	15	2,437	0,457	1,683
2004	0	0	0,032	0,015	56	13	2,611	1,503	1,484
2005	0	0,05	0,034	0,011	54	13	2,577	1,313	1,570
2006	0	0,02	0,032	0,012	46	15	2,486	0,831	1,702
2007	0	0,01	0,033	0,013	19	14	2,229	0,958	1,679
2008	0	0,06	0,037	0,013	22	14	2,258	0,829	1,705
2009	0	0,11	0,034	0,012	22	14	2,301	0,418	1,759
2010	0	0,04	0,027	0,015	15	14	2,335	-0,404	1,841
2011	0,01	0	0,031	0,019	23	6	2,533	1,781	1,262
2012	0	0	0,027	0,009	41	24	2,292	0,805	1,819
2013	0	0,04	0,028	0,009	42	26	2,298	0,618	1,858
2014	0	0,05	0,028	0,009	42	26	2,298	0,621	1,858

Table C.3: *P.values for the 9-digit data (window)*

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0,47	0,5	0,061	0,041	9	6	3,093	-0,058	1,278
1981	0,69	0,02	0,044	0,044	9	2	2,910	1,901	0,828
1982	0,19	0,18	0,061	0,033	11	3	2,886	2,005	0,878
1983	0,01	0,27	0,078	0,031	11	3	2,575	2,119	0,917
1984	0,6	0,9	0,064	0,018	25	3	3,207	2,152	0,982
1985	0,03	0,72	0,072	0,020	13	4	2,471	2,185	1,008
1986	0,2	0,91	0,073	0,018	27	4	3,020	2,235	1,021
1987	0,65	0,47	0,058	0,024	33	4	3,149	2,342	1,044
1988	0,43	0,6	0,070	0,023	37	7	3,164	2,442	1,023
1989	0	0,31	0,094	0,025	35	5	3,071	2,601	1,010
1990	0,34	0,23	0,066	0,026	40	2	3,274	2,407	1,169
1991	0,06	0,12	0,088	0,028	42	3	3,240	2,542	1,122
1992	0,08	0,49	0,091	0,021	62	3	3,696	2,599	1,115
1993	0,19	0,56	0,088	0,021	70	4	3,929	2,704	1,063
1994	0,15	0,53	0,089	0,020	70	2	3,914	2,644	1,123
1995	0	0,35	0,092	0,022	19	2	2,247	2,655	1,139
1996	0	0,36	0,090	0,022	40	2	2,795	2,707	1,132
1997	0	0,58	0,097	0,020	13	2	2,039	2,715	1,135
1998	0,01	0,25	0,094	0,023	52	2	3,127	2,756	1,142
1999	0	0,32	0,087	0,021	26	4	2,393	2,887	1,079
2000	0	0,2	0,088	0,026	34	8	2,586	3,039	1,011
2001	0	0,11	0,091	0,026	49	5	2,862	3,023	1,050
2002	0,28	0,11	0,089	0,025	96	4	4,203	3,006	1,093
2003	0,05	0,25	0,096	0,024	100	5	4,319	3,119	1,042
2004	0	0,45	0,090	0,021	33	7	2,429	3,186	1,021
2005	0	0,36	0,087	0,023	32	7	2,381	3,214	1,020
2006	0	0,35	0,087	0,023	35	8	2,424	3,322	0,984
2007	0	0,12	0,092	0,026	35	9	2,433	3,375	0,970
2008	0	0,22	0,088	0,026	37	15	2,450	3,555	0,891
2009	0	0,25	0,088	0,025	37	15	2,440	3,561	0,901
2010	0	0,4	0,088	0,024	41	15	2,515	3,594	0,893
2011	0,47	0,1	0,074	0,029	123	17	4,438	3,676	0,856
2012	0	0,17	0,082	0,027	40	17	2,465	3,681	0,857
2013	0	0,21	0,087	0,027	42	17	2,545	3,678	0,859
2014	0	0,19	0,086	0,027	42	17	2,544	3,678	0,859

Table C.4: *P.values for the 4-digit data (+1)*

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0,01	0	0,074	0,044	13	1	2,875	1,821	1,055
1981	0,02	0	0,063	0,030	13	1	2,884	1,898	0,990
1982	0,67	0	0,031	0,030	15	1	2,777	1,920	1,060
1983	0,36	0,04	0,038	0,020	18	1	2,741	1,971	1,069
1984	0,49	0,01	0,032	0,020	20	1	2,694	2,041	1,055
1985	0,6	0,03	0,030	0,019	23	4	2,686	1,763	1,209
1986	0,1	0,1	0,040	0,017	24	5	2,591	1,489	1,333
1987	0,16	0,03	0,035	0,019	21	5	2,527	1,724	1,318
1988	0,2	0,46	0,039	0,011	33	5	2,648	1,762	1,318
1989	0,63	0,44	0,030	0,011	35	4	2,724	2,019	1,240
1990	0,07	0,18	0,040	0,013	39	5	2,713	2,037	1,265
1991	0,31	0,71	0,037	0,009	44	4	2,760	2,133	1,247
1992	0,01	0,83	0,045	0,008	42	4	2,636	2,139	1,259
1993	0,01	0,54	0,045	0,011	43	11	2,616	1,793	1,372
1994	0	0,53	0,047	0,011	27	13	2,454	1,562	1,435
1995	0	0,26	0,044	0,013	26	12	2,403	1,677	1,413
1996	0,01	0,19	0,038	0,014	28	13	2,437	1,687	1,424
1997	0,18	0,17	0,044	0,013	92	13	3,033	1,712	1,422
1998	0	0,25	0,040	0,012	27	13	2,390	1,791	1,406
1999	0	0,15	0,041	0,014	28	14	2,387	1,665	1,451
2000	0	0,5	0,039	0,011	27	13	2,359	1,923	1,388
2001	0	0,1	0,046	0,013	34	11	2,419	2,164	1,323
2002	0	0,05	0,039	0,012	31	11	2,381	2,224	1,326
2003	0	0,03	0,042	0,011	33	5	2,394	2,484	1,245
2004	0	0,03	0,038	0,012	34	6	2,383	2,461	1,268
2005	0	0,05	0,041	0,012	36	5	2,383	2,523	1,258
2006	0	0,06	0,037	0,010	34	4	2,358	2,598	1,242
2007	0	0,47	0,038	0,010	37	14	2,359	2,123	1,423
2008	0	0,36	0,040	0,010	39	14	2,370	2,276	1,390
2009	0	0,3	0,043	0,010	40	14	2,356	2,308	1,395
2010	0	0,24	0,038	0,011	38	17	2,337	2,196	1,435
2011	0	0,17	0,036	0,011	41	15	2,361	2,293	1,414
2012	0	0,09	0,035	0,010	39	4	2,334	2,704	1,276
2013	0	0,27	0,036	0,010	39	14	2,335	2,374	1,396
2014	0	0,23	0,036	0,010	39	14	2,334	2,374	1,396

Table C.5: *P.values for the 7-digit data (+1)*

Year	pvalue.pl	pvalue.ln	KS.pl	KS.ln	xmin.pl	xmin.ln	para.pl	para.ln.1	para.ln.2
1980	0	0	0,085	0,031	12	1	2,695	2,255	0,874
1981	1	0	0,039	0,022	59	3	4,233	2,315	0,860
1982	0	0	0,057	0,017	25	1	2,983	2,281	0,944
1983	0	0	0,040	0,016	23	8	2,712	1,608	1,223
1984	0,09	0,08	0,031	0,014	28	8	2,764	1,589	1,249
1985	0,03	0,05	0,030	0,013	27	8	2,707	1,413	1,330
1986	0,04	0,23	0,027	0,011	27	10	2,654	0,942	1,466
1987	0,07	0,25	0,025	0,010	26	10	2,565	1,225	1,446
1988	0	0,03	0,030	0,013	26	11	2,537	1,258	1,454
1989	0,02	0	0,026	0,013	26	9	2,529	1,625	1,361
1990	0,01	0	0,027	0,012	35	8	2,641	1,894	1,310
1991	0,02	0	0,027	0,017	42	8	2,636	1,938	1,317
1992	0,01	0	0,023	0,016	35	21	2,587	1,030	1,540
1993	0,03	0,03	0,021	0,014	34	22	2,547	0,060	1,742
1994	0	0,07	0,025	0,012	41	22	2,551	0,052	1,759
1995	0,01	0,04	0,023	0,014	41	27	2,554	0,534	1,676
1996	0,01	0,1	0,023	0,012	45	30	2,561	1,269	1,539
1997	0	0,21	0,029	0,012	50	34	2,578	1,075	1,589
1998	0	0,51	0,030	0,010	51	33	2,574	1,008	1,607
1999	0	0,18	0,027	0,009	34	22	2,450	0,931	1,625
2000	0	0,27	0,027	0,009	34	24	2,436	0,353	1,752
2001	0	0,12	0,027	0,010	38	28	2,444	0,484	1,736
2002	0	0,17	0,026	0,009	29	23	2,379	0,410	1,763
2003	0	0,22	0,027	0,008	65	24	2,544	0,601	1,735
2004	0	0,35	0,028	0,008	65	26	2,527	0,511	1,768
2005	0	0,06	0,027	0,010	35	23	2,371	0,817	1,719
2006	0	0,4	0,026	0,007	39	26	2,376	0,195	1,856
2007	0	0,08	0,025	0,009	38	24	2,347	0,430	1,832
2008	0	0,21	0,026	0,008	39	29	2,337	0,398	1,853
2009	0	0,04	0,028	0,008	37	24	2,294	0,771	1,801
2010	0	0,1	0,028	0,008	37	24	2,285	0,780	1,810
2011	0	0,01	0,026	0,010	39	24	2,295	0,784	1,817
2012	0	0	0,027	0,009	41	24	2,292	0,805	1,819
2013	0	0,04	0,028	0,009	42	26	2,298	0,618	1,858
2014	0	0,05	0,028	0,009	42	26	2,298	0,621	1,858

Table C.6: *P.values for the 9-digit data (+1)*

Appendix D

Core-periphery network structure identification

Some networks are defined by a densely interconnected core and a more or less sparsely connected periphery as shown in Figure D.3. This type of structure has been identified in citation networks, the internet and lexicographical networks amongst others. This particular structure results in a core of a few densely connected networks and a periphery of many sparsely connected nodes. Having this type of structure makes for a particular degree distribution when compared to network with a more homogenous distribution. There should hence we a method to identify a core-periphery structure statistically. In order to identify the structure we start with plotting the cumulative degree distribution of a network. Figure D.1 gives an example of a degree distribution. This distribution gives the degree on the x-axis and the number of nodes with that degree on the y-axis. From this distribution we can see that the number of nodes with a high degree is low. In addition, the number of nodes with only a few links is high. This information alone is not sufficient to conclude that the network has a core-periphery structure. In order to get more precise information out of this data we are going to transform the degree distribution into a cumulative frequency distribution (Figure D.2).

The Cumulative Frequency Distribution (CFD) transforms the degree distribution into a probability distribution. From this distribution we can read the probability that a node taken at random from the graph had degree x. Note that this distribution is plotted in a log-log scale. The CFD then represents an equation linking frequency and degree. According to the network that is being represented the CFD highlights specific aspects of

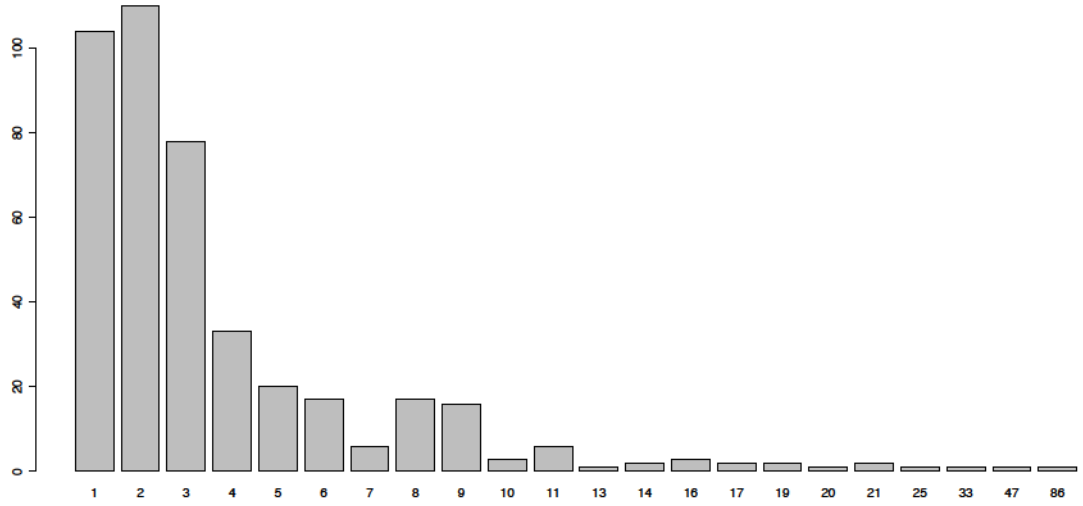


Figure D.1: *Example of a degree distribution*

the network structure. In this example we can see that the relation between the density and the frequency is linear. As such the relation can be written:

$$ln(y) = a \cdot ln(x) + b \quad \forall \ a < 0 \tag{D.1}$$

This is the equation for the log-log scale. On the more regular scale the form of this function is given by:

$$e^{ln(y)} = e^{a \cdot ln(x) + b} \tag{D.2}$$

$$y = e^{a \cdot ln(x)} \cdot e^{b} \tag{D.3}$$

$$y = e^{ln(x^a)} \cdot e^{b} \tag{D.4}$$

$$y = e^{b} \cdot x^{a} y = C \cdot x^{a} \tag{D.5}$$

This highlights the fact that when we increase the density by a factor of k, the frequency drops by a factor k^a with $a < 0$. The latter is true for each value the density might take. For this reason, when the CFD of a network has a linear form on the log-log scale, the network is referred to as a scale-free network.

The scale-free network is not the only core-periphery structure. Exponential and log-normal distribution can also represent core-periphery structures. The main difference

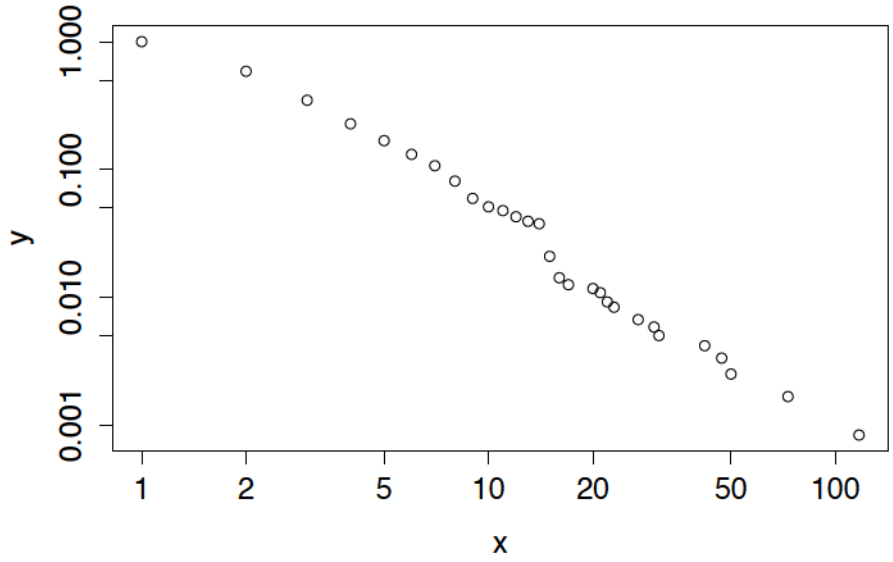


Figure D.2: *Example of a cumulative frequency distribution*

between the distribution is the manner in which the core transitions to the periphery. In a very abrupt case as in figure D.3 the transition is very abrupt. 67% of the nodes has a degree of one while the 33% of the nodes have a degree of 5. The CFD will show a sharp drop in frequency between densities 1 and 5. The scale-free structure is a particular cas in which the decrease in frequency is constant. Another case can be imagined in which there are many nodes of degree 1, 2, 3 and 4 making for a more dense periphery.

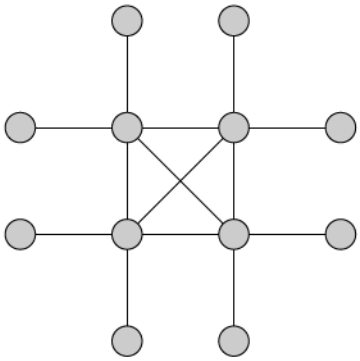


Figure D.3: *Core-Periphery illustration*

Appendix E

Initial parameters for the ABM model

Parameter	Interval
trajStep	0.2 - 0.4
β_0	0.05 - 0.2
ζ	0.01 - 0.05
δ_{rad}	0.1 - 0.5
knowledge depreciation	0.05 -0.2
ϕ	0.01 - 0.05
Window-productivity	10 - 20
weight	0.2 - 0.8
γ_{inc}	0.001 - 0.009
γ_{rad}	0.001 - 0.009
socialDimension.inc	0 - 1
socialDimension.rad	0 - 1
ProbImit	0.01 - 0.04

Table E.1: *Parameters and their intervals in the Monte-Carlo simulation*

Parameter/variable	Value
Initial Capital	100
Scrap rate	0.9
Cost	0.16
Depreciation	0.03
Interest rate	0.01
Floor productivity rate	0.05
Demand	300

Table E.2: *Fixed initial parameters and variables*

Appendix F

Gatekeepers in the collaboration network for Structural Composite Materials

Firm name	1996	2000	2004	2008	2014
AIRBUS	■	■	■	■	■
BOEING	■	■	■	■	■
GE	■	■	■	■	■
NASA	■	■	■	■	■
NORTHWESTERN UNIV	■	■	■	■	■
PECHINEY	■	■	■	■	■
SICMA	■	■	■	■	■
UNIV OF DELAWARE	■	■	■	■	■
UNIV OF VIRGINIA	■	■	■	■	■
ASTRIUM		■	■	■	■
DASSAULT		■	■	■	■
EUROCOPTER		■	■	■	■
HONEYWELL		■	■	■	■
ROLLS ROYCE		■	■	■	■
SAINT GOBAIN		■	■	■	■
BAE SYSTEMS			■	■	■
CNES			■	■	■
EADS			■	■	■
FRAUNHOFER			■	■	■
INDUSTRIELLE DU PONANT			■	■	■
MITSUBISHI EIT			■	■	■
NIPPON STEEL CORP			■	■	■
ALSTOM				■	■
DAHER SOCATA				■	■
EUROP PROPULSION				■	■
MESSIER BUGATTI DOWTY				■	■
RENAULT				■	■
SNECMA				■	■
TECHSPACE				■	■
UNIV ORLEANS				■	■
CNRS					■
HISPANO HUREL					■
UNIV LORRAINE					■
UNIV REIMS					■

Table F.1: *Gatekeepers between the publication and patent collaboration networks*

Appendix G

ERGM algorithm output

G.1 Stepping algorithm output

```
Iteration # 1 .   Trying gamma= 0.17
Iteration # 2 .   Trying gamma= 0.14
Iteration # 3 .   Trying gamma= 0.16
Iteration # 4 .   Trying gamma= 0.22
Iteration # 5 .   Trying gamma= 0.25
Iteration # 6 .   Trying gamma= 0.34
Iteration # 7 .   Trying gamma= 0.31
Iteration # 8 .   Trying gamma= 0.46
Iteration # 9 .   Trying gamma= 0.74
Iteration # 10 .   Trying gamma= 0.97
Iteration # 11 .   Trying gamma= 1
Iteration # 12 .   Trying gamma= 1
Now ending with one large sample for MLE.
Evaluating log-likelihood at the estimate. Using 20 bridges: 1 2 3 4 5
   6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 .
```

Figure G.1: *R output for the Stepping algorithm*

G.2 Robbins-Monro algorithm output

```
Robbins-Monro algorithm with theta_0 equal to:
    edges  triangle
-4.676219  1.456380
Phase 1:  13 iterations (interval=1024)
Phase 1 complete; estimated variances are:
   edges triangle
3676.692 1175.308
Phase 2, subphase 1 : a= 0.1 , 9 iterations (burnin=16384)
theta new: -4.66068075119643
 theta new: 1.42768924899568
Phase 2, subphase 2 : a= 0.05 , 23 iterations (burnin=16384)
theta new: -4.64740903958273
 theta new: 1.4235669242362
Phase 2, subphase 3 : a= 0.025 , 58 iterations (burnin=16384)
theta new: -4.62881856406474
 theta new: 1.41405593966042
Phase 2, subphase 4 : a= 0.0125 , 146 iterations (burnin=16384)
theta new: -4.60985096388914
 theta new: 1.39932813390954
Phase 3:  20 iterations (interval=1024)
Evaluating log-likelihood at the estimate.
```

Figure G.2: *R output for the Robbins-Monro algorithm*

Appendix H

Introduction (Version Française)

Le paysage technologique est en constante évolution. De nouvelles technologies sont découvertes, d'autres sont recombinées pour créer un nouveau produit. Dans ce context de changement technologique, les firmes doivent s'adapter pour survivre sur le marché. Pour s'assurer de ne pas perdre des parts de marché, une firme doit s'assurer de maîtriser toutes les technologies pour faire évoluer leurs produits. Cependant, avec la complexité croissante des produits, le nombre de technologies que la firme doit maîtriser augmente. Il se peut que la firme se heurte au problème de ne pas maîtriser les technologies requises pour continuer le développement de son produit. Lorsque la firme est confrontée à un manque de savoir-faire elle est confrontée au choix de rechercher et développer la nouvelle technologie par ses propres moyens ou de chercher un collaborateur qui détient déjà des compétences en la matière. Tenter de développer la technologie soi-même est un choix risqué qui demande de gros investissements en termes financier et en termes de temps. La collaboration apparaît alors comme une solution viable pour innover.

Depuis plusieurs décennies, le nombre de collaborations n'a cessé de croître (Saviotti, 2007; Tomasello et al., 2013). Des collaborations entre universités et firmes, entre fournisseurs et donneur d'ordre et même entre concurrents, deviennent un phénomène courant. La collaboration permet aux firmes de mettre en commun leurs connaissances et capital productif pour innover. Ce point est d'autant plus important que les technologies qui sont développées par la firme sont complexes ou diverses.

Ces collaborations permettent aux firmes d'échanger des connaissances. Ces échanges peuvent être volontaires (échange de technologie, accord de licence, formation) ou involontaires (on observe le fonctionnement à de l'autre firme et on imite) pensez à une meilleure

organisation, optimisation de la configuration des machines, méthodes managériales etc. Les collaborations permettent alors à la fois la création de nouvelles connaissances et la diffusion de ces dernières entre partenaires. Cet échange de connaissance n'est cependant pas parfait. C'est à la firme de décider des connaissances qu'elle est prête à mettre à la disposition de ses collaborateurs. De plus, la qualité de ce transfert dépend de la capacité de la firme à envoyer ou à absorber les connaissances auxquelles elle est exposée. Une firme qui est exposée à des connaissances qui sont trop avancées pour elle, ne pourra les intégrer dans son processus de R&D. Une firme exposée à des connaissances que la firme maîtrise déjà trouvera que ces connaissances n'ont pas un impact significatif sur son processus de R&D.

Du côté du transfert Il se peut aussi firme ne soit pas apte, ou mette de la mauvaise volonté dans la transmission des connaissances. Les firmes peuvent donc inhiber ou catalyser le transfert des connaissances. Ces éléments prennent toute leur importance lorsque l'on regarde ce échanges d'un œil plus méso-macro. Les firmes peuvent en effet avoir plus qu'un collaborateur, qui à leur tour peuvent avoir des collaborateurs et ainsi de suite. Les connaissances détenues par les firmes, combinées avec celles reçues peuvent êtres transmises à d'autres collaborateurs. De firme à firme les connaissances diffusent à travers le réseau des collaborateurs, soit accélérées soit ralenties par les firmes. La manière dont ces firmes sont interconnectées, la structure du réseau, définit en partie la vitesse avec laquelle les connaissances diffusent entre les firmes. Idéalement, on souhaiterait que toutes les connaissances soient à la disposition de tout le monde, mais cette vision est bien trop utopique. Dans un réseau qui soit densément interconnecté, les connaissances peuvent être envoyées par différents chemins simultanément. De plus, la distance faible qui sépare les firmes dans le réseau fait que les connaissances sont rapidement accessibles. Dans le cas contraire, dans un réseau peu dense le temps de diffusion peut-être fortement accru car le nombre de chemins reliant deux firmes est plus faible. Si une firme bloque le transfert, il est difficile de récupérer les connaissances par un autre chemin. Dans ce type de structure des firmes avec une position particulière, les « gatekeepers » peuvent faire leur apparition. Les « gatekeepers » sont des firmes qui ont la particularité de se trouver sur un chemin unique entre plusieurs clusters de firmes. Elle a donc accès à des connaissances spécifiques venant d'un des clusters et peut décider si elle souhaite partager ces connaissances avec un autre cluster. Ceci fait des gatekeepers des firmes importantes pour la diffusion des

connaissances. Elles peuvent bloquer la diffusion si la firme est inefficace dans la transmission. La structure du réseau est donc une variable d'intérêt lorsque l'on étudie les échanges de connaissances. Ces aspects soulèvent deux premières questions que j'aborde dans cette thèse :

1. Existe-t-il une structure de réseau qui soit plus efficace qu'une autre d'un point de vue performance ?

2. Dans ces réseaux, existe-t-il des positions qui favorisent la performance des firmes vis-à-vis d'autres firmes ?

La diffusion des connaissances n'est pas la seule raison pour laquelle la structure du réseau est une variable d'intérêt. La structure est aussi importante car elle nous permet de mieux comprendre les stratégies de R&D des firmes qui la composent. Le réseau est une agrégation de collaborations qui sont la résultante de décisions stratégiques des firmes. La position de la firme dans le réseau permet de voir combien la firme a de collaborations, si elle collabore uniquement avec des firmes de son secteur, si elle s'est lancée dans des collaborations sur de nouveaux marchés, si elle a la position de gatekeeper sur certaines technologies etc. Si on regarde la structure globale du réseau on peut mieux comprendre les stratégies d'un secteur d'activité ou d'une région (en fonction de l'objectif du réseau) : est-ce que les innovations sont créées localement dans le réseau (dans des clusters), quel est la place des institutions de recherche, il y a-t-il des gatekeepers, est-ce que les compétiteurs collaborent, dans quelle phase du cycle de vie de la technologie se trouve la technologie. Comprendre la structure du réseau revient à avoir une vision du processus de R&D sur une plus grande échelle.

La dernière question à laquelle cette thèse tente de répondre est : 3. Comment analyser et interpréter la structure d'un réseau de collaboration en termes des stratégies de R&D ?

Bien sûr, ceci varie d'un secteur à un autre et avec l'échelle d'analyse. Le réseau de collaboration autour d'une technologie précise n'évolue pas de la même manière qu'un réseau de collaboration au niveau d'un secteur ou d'une région géographique. Le réseau au niveau du secteur est une coexistence de collaborations autour de technologies qui évoluent à différents stades d'avancement. Les stratégies de R&D relatives à la technologie sont donc difficilement observables au niveau sectoriel et régional. Il est donc important d'analyser un réseau de collaboration dans son contexte. Pour cette raison je propose dans cette thèse des analyses empiriques se focalisant sur différents contextes.

L'objectif de cette première analyse est d'étudier un des facteurs qui pourrait intervenir dans la dynamique des réseaux de collaboration : le cycle de vie de la technologie. L'étude du cycle de vie est importante car elle permet aux firmes de savoir à quel moment entrer sur un marché ou à quel moment il faut commencer à se tourner vers une nouvelle technologie. Le réseau de collaboration permet alors de voir quels sont les acteurs présents dans le domaine et quels seraient de potentiels collaborateurs.

Sachant que le cycle de vie de la technologie est facteur qui est inobservable à un niveau plus agrégé, ce chapitre est focalisé sur une technologie en particulier : les composites structuraux en aéronautique. Le développement d'une technologique est accompli en deux grandes étapes, une phase de recherche de la technologie suivie d'une seconde phase, la phase développement. Pendant chacune de ces phases la firme n'a pas les mêmes collaborateurs. La phase développement demande de nouvelles collaborations pour intégrer la technologie recherché pour de nouvelles applications. La structuration du réseau devrait donc être modifié lors du changement de phase. Notons ici que cette idée ne remet aucunement en cause la vision plus récente du cycle de vie des technologies basé sur le feedback. Les firmes peuvent continuer à collaborer avec les mêmes firmes que lors de la phase recherche, cela n'a simplement pas de coïncidence sur la structure du réseau de collaboration (sauf sur le poids des liens). Pour identifier le cycle de vie de la technologie, ce chapitre introduit une nouvelle méthode basée sur une analyse réseau des codes CIB (Classement International des Brevets). Sur chaque brevet se trouvent des codes CIB qui décrivent la technologie contenue dans le brevet. Ainsi, un brevet peut contenir plusieurs codes, témoignant de la recombinaison des technologies qui ont menées à produire la technologie qui fait l'objet du brevet. En créant un réseau à partir de ces codes on s'attend à voir apparaître un réseau dense, composé des codes qui sont à la base de la technologie, ou le cœur de la technologie. En effet, tous les brevets déposés sur le cœur de la technologie contiennent majoritairement les mêmes codes. Une fois la technologie prête pour être exploitable, la phase développement commence. On voit alors apparaître des brevets qui contiennent des codes qui font référence à la technologie fondamentale et des codes qui font référence à l'application. Prenons par exemple la photographie et son application dans les Smartphones. On a un cœur technologique spécifique à la photographie constitué de tous les brevets déposés sur la photographie. L'application de cette technologie dans les Smartphones fait l'objet de nouveaux brevets contenant des codes relatifs aux Smartphones

et à la photographie. On a donc des codes qui viennent se greffer au réseau dense, créant une périphérie autour du cœur.
Si on observe une variation dans la structure du réseau de collaboration autour de la période où la technologie change de cycle on pourra conclure que le cycle de vie de la technologie a un impact significatif sur la formation du réseau de collaboration et permettra donc de mieux comprendre l'évolution des réseaux de collaboration.

Les analyses qui suivent ce premier chapitre approfondissent l'analyse de la structure du réseau de collaboration. Ces deux chapitres changent d'optique et se focalisent sur le réseau de collaboration de deux secteurs en France : le secteur aéronautique et le secteur des biotechnologies. Ces deux secteurs ont été retenus pour leurs différences structurelles notables, le premier est organisé en chaine de valeur alors que le second est un secteur atomisé hautement concurrentiel. Ces caractéristiques devraient jouer sur la structure du réseau rendant une comparaison intéressante.
L'objectif de ces deux analyses est double. Une première partie de ces analyses se focalise sur la structure du réseau et sa dynamique. La seconde cherche à identifier un lien entre la position de la firme dans le réseau et sa performance.
L'analyse de la structure se fait à trois niveaux, chacun apportant une information précise sur la structure. Au niveau le plus agrégé on analyse la structure globale du réseau. L'objectif de cette première étape est d'avoir un aperçu de la stratégie de R&D du secteur. Au delà de l'identification des firmes centrales et de la présence des institutions de recherche, je commence par vérifier si la structure globale présente des caractéristiques spécifique qui sont bien comprises par la littérature. Lorsqu'une structure présente ces caractéristiques, l'organisation du réseau devient plus claire. On peut facilement déduire si les acteurs sont proches ou éloignées, si le réseau a une forte tendance à se clustériser, vérifier la présence de triangles, vérifier si le réseau est homogène ou irrégulier, et si le réseau est irrégulier qui est favorisé.
On verra dans le chapitre 3 que lorsqu'un réseau ne présente pas de caractéristiques connues les choses se compliquent. Les structures pour lesquels je teste sont la structure petit-monde est la structure Scale-free. Les petits mondes sont observés majoritairement dans les réseaux sociaux. Ils sont identifiables par une distance moyenne faible entre les individus et une tendance forte pour le clustering. Le clustering est un indicateur réseau

qui mesure le nombre de triangles. Les triangles reflètent le fait que les amis de mes amis ont tendance à être mes amis. Ce principe est courant dans les réseaux sociaux mais a aussi un sens dans le cadre des collaborations. Une collaboration est une entreprise risquée, les recommandations jouent un rôle important dans le choix des collaborateurs car elles permettent de réduire le risque inné à la collaboration (risque de défaut, passager clandestin, mésentente). Les firmes peuvent demander des recommandations à leurs collaborateurs qui ne peuvent juger que de leurs propres collaborateurs. Le résultat est une collaboration entre collaborateurs qui donne un triangle au niveau du réseau, résultent en l'apparition de clusters localisés. Un petit monde est donc une structure dans laquelle les connaissances sont générées localement grâce aux clusters et qui diffuse rapidement partout dans le réseau grâce à la faible distance qui sépare les firmes.

La structure scale-free a des caractéristiques bien différentes. Cette structure est lié à la loi de Pareto (aussi connu sous le nom de loi 20-80).). Une faible portion des firmes possède un grand nombre de lien et une grande portion des firmes détient un nombre faible de liens. Ceci donne lieu a un cœur qui interconnecte les firmes avec beaucoup de liens et une périphérie formée des firmes avec peu de liens qui se connectent autour. Ces structures ont été observées dans les réseaux économiques mais aussi dans le réseau internet et les réseaux de citation. Cette structure nous informe donc que le réseau est construit à partir de certaines grandes firmes interconnectées à des firmes plus petites. Cette structure est donc susceptible d'apparaître dans le cas d'une chaine de valeur. Les grandes firmes, les assembleurs, sont connectées à l'ensemble de leurs sous-traitants. Les sous-traitants ont en revanche que très peu de liens. On s'attendrait donc à ce que le réseau aéronautique présente ces caractéristiques. Dans le cas du réseau des biotechnologies on s'attendrait plus à une structure de type petit-monde. La structure globale du réseau est un reflet des caractéristiques de l'objet d'étude sous-jacent.

La structure globale est composée de l'interconnexion de sous-réseaux, aussi appelés clusters. La seconde étape de l'analyse est donc de regarder de plus près les clusters qui composent la structure globale. Cette étape, que l'on peut qualifier d'analyse méso, est une étape cruciale dans la compréhension du processus de R&D. La présence (ou l'absence) de clusters permet de juger du caractère locale de la création des connaissances. Si le réseau est une interconnexion de clusters, les connaissances sont générées dans chaque cluster avant de diffuser. En l'absence de clusters les connaissances sont générées en mobilisant

le réseau dans son intégralité. J'utilise la méthode de Louvain pour identifier les clusters. Cette méthode maximise le nombre de liens dans la communauté tout en minimisant le nombre de liens entre communautés pour identifier les communautés dans les réseaux. Dans les analyses de cette thèse cette méthode donne des communautés bien définies et significatives.

La question qui reste sans réponse est alors de savoir comment ces communautés se sont formées. Pour répondre à cette question on passe à la dernière étape de l'analyse réseau : le niveau micro. Dans cette dernière étape on identifie les facteurs qui motivent une firme à collaborer avec une firme précise plutôt qu'une autre. Pour cela j'utilise une méthode qui pour l'instant n'est que peu utilisé en sciences économiques, Exponential Random Graph Models (ERGM). Ces modèles sont des régressions logistiques modifiées. Dans les régressions classiques on travaille avec l'hypothèse que les observations dont on dispose sont indépendantes. Cette hypothèse pose des problèmes lorsque l'on cherche analyse des réseaux. En effet, bien souvent la probabilité d'apparition d'un lien dans un réseau est dépendant de la structure du réseau. Par exemple, si une firme a deux collaborateurs, la probabilité qu'elles finissent par collaborer est plus élevée que pour deux firmes qui n'ont aucun collaborateur en commun. La probabilité d'apparition d'un lien peut donc dépendre de la structure du réseau avant la création de lien. Les modèles ERGM sont capables de prendre en compte ces dépendances et sont donc adaptés pour l'analyse des réseaux de collaboration. Sachant que cette méthode n'est que peu répandue un chapitre introductif à cette méthode ainsi qu'une application est présenté avant les analyses sectoriels.

Une fois que l'analyse de la structure du réseau est terminée j'aborde la question de la performance. En fonction de la position que les firmes ont dans le réseau elles n'ont pas accès aux mêmes connaissances, soit à cause de leur position structurelle donnant lieu a plus de trafic ou simplement car elles évoluent dans un voisinage plus diversifié. Lorsque j'analyse la performance de la firme je prends donc en compte à la fois les éléments structurels et les éléments relatifs au voisinage. La mesure de la performance retenue est la Return On Assets (ROA). Cette dernière a été retenue car elle représente la performance de la firme au sens large, elle inclut la propriété intellectuelle. Une régression panel est utilisée pour identifier les différents facteurs qui ont un impact significatif sur la performance de la firme.

Identifier la performance du réseau entier, la dernière question, est plus délicat à aborder

avec des données empiriques. En tout cas, les données dont je dispose ne me permettent pas de répondre à cette question de manière adéquate. Pour cette raison j'aborde cette question par un modèle théorique. La méthode de modélisation retenue est la modélisation à base d'agents. Sachant que la valeur d'un réseau de collaboration réside dans la diversité des connaissances et des agents qui la composent cette méthode paraît le choix naturel. Pour modéliser correctement le fonctionnement du réseau, le modèle doit inclure des mécanismes relatifs aux échanges de connaissances mais aussi relatifs à la manière dont la firme transforme ces connaissances en parts de marché. Le modèle ne doit donc pas se limiter à un modèle de diffusion des connaissances. Pour pouvoir mesurer la performance du réseau des mesures tels que le niveau technologique des firmes, le surplus des consommateurs ou encore le profit des firmes sont nécessaires. Dans cette thèse je pars d'un modèle existant (Jonard and Yildizoglu, 1999) qui inclut déjà certains de ces mécanismes en l'étendant de plusieurs manières. Une première extension consiste en l'inclusion d'un mécanisme d'innovation radicale qui permet aux firmes de changer de trajectoire technologique. Ensuite, les mécanismes régissant les échanges de connaissances et leur inclusion dans le processus de R&D ont été rendus plus explicites.

Il s'agit d'un modèle dans lequel « n » firmes évoluent dans un réseau fixé de manière exogène. Les firmes produisent un bien homogène avec une technologie donnée. En innovant, les firmes peuvent améliorer leur technologie de production et ainsi devenir plus efficace. Pour innover, les firmes utilisent leurs connaissances propres mais aussi les flux de connaissances auxquels elles sont exposés. Les firmes produisent un bien homogène avec une technologie donnée. En innovant, les firmes peuvent améliorer leur technologie de production et ainsi devenir plus efficace. Pour innover, les firmes utilisent leurs connaissances propres mais aussi les flux de connaissances auxquels elles sont exposées. En disposant les firmes sur différentes formes de réseau, et en faisant tourner les mêmes simulations un grand nombre de fois, on tire des conclusions sur la performance d'une structure par rapport à une autre.

Les différents éléments que je viens d'exposer font apparaître la structure de la thèse. Un premier chapitre survole la littérature pour identifier les grandes questions qui seront abordées dans la suite de la thèse. Une première analyse se focalisant sur l'influence du cycle de vie de la technologie comme facteur explicatif de la structuration du réseau est alors proposée. Avant de passer aux analyses au niveau sectoriel, les ERGM sont

introduits. Les deux chapitres qui suivent visent à identifier les facteurs explicatifs de la structure du réseau dans un premier temps et analysent l'impact de la performance dans un second temps. Le dernier chapitre propose un modèle théorique qui analyse l'efficacité de différents structures de réseaux.

Bibliography

Bibliography

Ahlström-Söderling, R.: 2003, Sme strategic business networks seen as learning organizations, *Journal of Small Business and Enterprise Development* **10**(4), 444–454.

Ahn, Y.-Y., Bagrow, J. P. and Lehmann, S.: 2010, Link communities reveal multiscale complexity in networks, *Nature* **466**(7307), 761–764.

Ahuja, G.: 2000, Collaboration networks, structural holes, and innovation: A longitudinal study, *Administrative science quarterly* **45**(3), 425–455.

Alencar, M., Porter, A. and Antunes, A.: 2007, Nanopatenting patterns in relation to product life cycle, *Technological Forecasting and Social Change* **74**(9), 1661–1680.

Alghamdi, M., McDonald, S. and Pailthorpe, B.: 2012, *The Emergence of a Small World in a Network of Research Joint Ventures*, University of Queensland, School of Economics.

Anderson, C. J., Wasserman, S. and Crouch, B.: 1999, A p* primer: Logit models for social networks, *Social Networks* **21**(1), 37–66.

Arrow, K. J.: 1962, Economic welfare and the allocation of resources for invention, *The RAND Corporation* pp. 609–626.

Bala, V. and Goyal, S.: 2000, A noncooperative model of network formation, *Econometrica* **68**(5), 1181–1229.

Bar, T. and Leiponen, A.: 2012, A measure of technological distance, *Economics Letters* **116**(3), 457–459.

Barabási, A.-L. and Albert, R.: 1999, Emergence of scaling in random networks, *science* **286**(5439), 509–512.

Barringer, B. R. and Harrison, J. S.: 2000, Walking a tightrope: Creating value through interorganizational relationships, *Journal of management* **26**(3), 367–403.

Bastian, M., Heymann, S. and Jacomy, M.: 2009, Gephi: An open source software for exploring and manipulating networks.

Baum, J. A., Shipilov, A. V. and Rowley, T. J.: 2003, Where do small worlds come from?, *Industrial and Corporate change* **12**(4), 697–725.

Belke, A. and Polleit, T.: 2006, Money and swedish inflation, *Journal of Policy Modeling* **28**(8), 931–942.

Belongia, M. and Ireland, P.: 2014, A working solution to the question of nominal gdp targeting, *Macroeconomic Dynamics* **31**. cited By 0.

Besag, J. E.: 1972, Nearest-neighbour systems and the auto-logistic model for binary data, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 75–83.

Besag, J. E.: 1975, Statistical analysis of non-lattice data, *The statistician* pp. 179–195.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E.: 2008, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment* **2008**(10), P10008.

Bloom, N., Schankerman, M. and Van Reenen, J.: 2013, Identifying technology spillovers and product market rivalry, *Econometrica* **81**(4), 1347–1393.

Breschi, S. and Lissoni, F.: 2001, Knowledge spillovers and local innovation systems: a critical survey, *Industrial and corporate change* **10**(4), 975–1005.

Breschi, S., Lissoni, F. and Malerba, F.: 2003, Knowledge-relatedness in firm technological diversification, *Research Policy* **32**(1), 69–87.

Broekel, T. and Hartog, M.: 2013, Explaining the structure of inter-organizational networks using exponential random graph models, *Industry and Innovation* **20**(3), 277–295.

Buchmann, T. and Pyka, A.: 2013, The evolution of innovation networks: The case of a german automotive network, *Technical report*, FZID discussion papers.

Burt, R. S.: 2004, Structural holes and good ideas1, *American journal of sociology* **110**(2), 349–399.

Butts, C. T.: 2008, network: a package for managing relational data in r., *Journal of Statistical Software* **24**(2).

Caimo, A. and Lomi, A.: 2015, Knowledge sharing in organizations: A bayesian analysis of the role of reciprocity and formal structure, *Journal of Management* **41**(2), 665–691. cited By 2.

Camisón, C. and Forés, B.: 2011, Knowledge creation and absorptive capacity: The effect of intra-district shared competences, *Scandinavian Journal of Management* **27**(1), 66–86.

Carayol, N., Roux, P. and Yıldızoğlu, M.: 2008, In search of efficient network structures: the needle in the haystack, *Review of Economic Design* **11**(4), 339–359.

Carrington, P. J., Scott, J. and Wasserman, S.: 2005, *Models and methods in social network analysis*, Vol. 28, Cambridge university press.

Chang, S.-B.: 2012, Using patent analysis to establish technological position: Two different strategic approaches, *Technological Forecasting and Social Change* **79**(1), 3–15.

Chen, Y., Lin, M. and Chang, C.: 2009, The positive effects of relationship learning and absorptive capacity on innovation performance and competitive advantage in industrial markets, *Industrial Marketing Management* **38**(2), 152–158.

Cohen, W. M. and Levin, R. C.: 1989, Empirical studies of innovation and market structure, *Handbook of industrial organization* **2**, 1059–1107.

Cohen, W. M. and Levinthal, D. A.: 1990, Absorptive capacity: A new perspective on learning and innovation, *Administrative Science Quarterly* **35**(1), 128–152.

Comin, D. and Hobijn, B.: 2004, Cross-country technology adoption: making the theories face the facts, *Journal of monetary Economics* **51**(1), 39–83.

Cowan, R. and Jonard, N.: 2004, Network structure and the diffusion of knowledge, *Journal of economic Dynamics and Control* **28**(8), 1557–1575.

Cowan, R. and Jonard, N.: 2007, Structural holes, innovation and the distribution of ideas, *Journal of Economic Interaction and Coordination* **2**(2), 93–110.

Cranmer, S., Desmarais, B. and Menninga, E.: 2012, Complex dependencies in the alliance network, *Conflict Management and Peace Science* **29**(3), 279–313. cited By 13.

Czudaj, R.: 2011, P-star in times of crisis - forecasting inflation for the euro area, *Economic Systems* **35**(3), 390–407.

Davide Chiaroni, V. C. . F. F.: 2008, Patterns of collaboration along the biopharmaceutical innovation process, *Journal of Business Chemistry* **5**(1), 7–22.

Dosi, G.: 1982, Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change, *Research policy* **11**(3), 147–162.

Dosi, G.: 2000, *Innovation, organization and economic dynamics: selected essays*, Edward Elgar Publishing.

Duanmu, J.-L. and Fai, F. M.: 2007, A processual analysis of knowledge transfer: From foreign mnes to chinese suppliers, *International Business Review* **16**(4), 449–473.

Duysters, G., Kok, G. and Vaandrager, M.: 1999, Crafting successful strategic technology partnerships, *R&D Management* **29**(4), 343–351.

Efron, B.: 1975, Defining the curvature of a statistical problem (with applications to second order efficiency), *The Annals of Statistics* pp. 1189–1242.

Egbetokun, A. A. and Savin, I.: 2013, Emergence of innovation networks from r&d cooperation with endogenous absorptive capacity, *Technical report*, Jena Economic Research Papers.

Frank, O. and Strauss, D.: 1986, Markov graphs, *Journal of the american Statistical association* **81**(395), 832–842.

Frigant, V., Kechidi, M. and Talbot, D.: 2006, Les territoires de l'aéronautique: Eads, entre mondialisation et ancrage.

Gao, L., Porter, A. L., Wang, J., Fang, S., Zhang, X., Ma, T., Wang, W. and Huang, L.: 2013, Technology life cycle analysis method based on patent documents, *Technological Forecasting and Social Change* **80**(3), 398–407.

Gay, B. and Dousset, B.: 2005, Innovation and network structural dynamics: Study of the alliance network of a major sector of the biotechnology industry, *Research policy* **34**(10), 1457–1475.

Geenhuizen, M. V.: 2008, Knowledge networks of young innovators in the urban economy: biotechnology as a case study, *Entrepreneurship and Regional Development* **20**(2), 161–183.

Geyer, C. J. and Thompson, E. A.: 1992, Constrained monte carlo maximum likelihood for dependent data, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 657–699.

Giuliania, E. and Bella, M.: 2005, The micro-determinants of meso-level learning and innovation: evidence from a chilean wine cluster, *Research Policy* **34**, 47–68.

Goyal, S. and Joshi, S.: 2003, Networks of collaboration in oligopoly, *Games and Economic behavior* **43**(1), 57–85.

Goyal, S. and Moraga-Gonzalez, J. L.: 2001, R&d networks, *Rand Journal of Economics* pp. 686–707.

Granovetter, M.: 1973, The strength of weak ties, *American journal of sociology* pp. 1360–1380.

Guida, M. and Maria, F.: 2007, Topology of the italian airport network: A scale-free small-world network with a fractal structure?, *Chaos, Solitons & Fractals* **31**(3), 527–536.

Gulati, R.: 1995, Social structure and alliance formation patterns a longitudianl analysis, *Administrative Science Quaterly* **40**, 619–652.

Gulati, R., Lavie, D. and Madhavan, R. R.: 2011, How do networks matter? the performance effects of interorganizational networks, *Research in Organizational Behavior* **31**, 207–224.

Gulati, R., Sytch, M. and Tatarynowicz, A.: 2012, The rise and fall of small worlds: Exploring the dynamics of social structure, *Organization Science* **23**(2), 449–471.

Hagedoorn, J.: 1996, Trends and patterns in strategic technology partnering since the early seventies, *Review of industrial Organization* **11**(5), 601–616.

Hagedoorn, J.: 2002, Inter-firm r&d partnerships: an overview of major trends and patterns since 1960, *Research policy* **31**(4), 477–492.

Hagedoorn, J. and Narula, R.: 1996, Choosing organizational modes of strategic technology partnering: international and sectoral differences, *Journal of international business studies* **27**(2), 265–284.

Hammersley, J. M. and Clifford, P.: 1971, Markov fields on finite graphs and lattices.

Hanaki, N., Nakajima, R. and Ogura, Y.: 2010, The dynamics of r&d network in the it industry, *Research policy* **39**(3), 386–399.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., Bender-deMoll, S. and Morris, M.: 2008, statnet: Software tools for statistical analysis of network data, *Journal of Statistical Software* **24**(1), 1 – 11.

Handcock, M. S., Robins, G., Snijders, T. A., Moody, J. and Besag, J.: 2003, Assessing degeneracy in statistical models of social networks, *Technical report*, Citeseer.

Hansen, M. T.: 1999, The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits, *Administrative science quarterly* **44**(1), 82–111.

Hargadon, A. B.: 2002, Brokering knowledge: Linking learning and innovation, *Research in Organizational behavior* **24**, 41–85.

Harris, J. K.: 2013, *An introduction to exponential random graph modeling*, Vol. 173, Sage Publications.

Herings, P., Mauleon, A. and Vannetelbosch, V. J.: 2014, Stability of networks under level-k farsightedness, *Ana and Vannetelbosch, Vincent J., Stability of Networks Under Level-K Farsightedness (July 10, 2014)* .

Human, S. E. and Provan, K. G.: 1997, An emergent theory og structure and outcomes in small-firm strategic manufacturing networks, *The Academy of Management Journal* **40**(2), 368–403.

Hummel, R. M., Hunter, D. R. and Handcock, M. S.: 2012, Improving simulation-based algorithms for fitting ergms, *Journal of Computational and Graphical Statistics* **21**(4), 920–939.

Hunter, D. R.: 2007, Curved exponential family models for social networks, *Social networks* **29**(2), 216–230.

Hunter, D. R. and Handcock, M. S.: 2006, Inference in curved exponential family models for networks, *Journal of Computational and Graphical Statistics* **15**(3).

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M.: 2008, ergm: A package to fit, simulate and diagnose exponential-family models for networks, *Journal of statistical software* **24**(3), nihpa54860.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., Morris and Martina: 2008, ergm: A package to fit, simulate and diagnosose exponential-family models for networks, *Journal of Statistical Software* **24**(3), 1 – 29.

Inventions, G.: 2002, Intellectual property rights and licensing practices.

Jackson, M. O.: 2003, The stability and efficiency of economic and social networks, *Networks and Groups*, Springer, pp. 99–140.

Jackson, M. O.: 2005, The economics of social networks.

Jackson, M. O. and Wolinsky, A.: 1996, A strategic model of social and economic networks, *Journal of economic theory* **71**(1), 44–74.

Jackson, M. O. and Yariv, L.: 2007, Diffusion of behavior and equilibrium properties in network games, *The American economic review* **97**(2), 92–98.

Jaffe, A. B.: 1986, Technological opportunity and spillovers of r&d: evidence from firms' patents, profits and market value.

Jonard, N. and Yildizoglu, M.: 1999, *Sources of technological diversity*, CNRS.

Kalinka, A. T. and Tomancak, P.: 2011, linkcomm: an r package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type, *Bioinformatics* **27**(14).

Kogut, B.: 2000, The network as knowledge: Generative rules and the emergence of structure, *Strategic management journal* **21**(3), 405–425.

Kogut, B. and Zander, U.: 1992, Knowledge of the firm, combinative capabilities, and the replication of technology, *Organization science* **3**(3), 383–397.

König, M. D., Battiston, S., Napoletano, M. and Schweitzer, F.: 2012, The efficiency and stability of r&d networks, *Games and Economic Behavior* **75**(2), 694–713.

König, M., Tessone, C. J. and Zenou, Y.: 2009, A dynamic model of network formation with strategic interactions.

König, Michael D. & Battiston, S. . N. M. . S. F.: 2011, Recombinant knowledge and the evolution of innovation networks, *Journal of Economic Behavior & Organization* **79**(3), 145–164.

Leung, R. C.: 2013, Networks as sponges: International collaboration for developing nanomedicine in china, *Research Policy* **42**(1), 211–219.

Lomi, A. and Fonti, F.: 2012, Networks in markets and the propensity of companies to collaborate: An empirical test of three mechanisms, *Economics Letters* **114**(2), 216–220.

Lomi, A. and Pallotti, F.: 2012, Relational collaboration among spatial multipoint competitors, *Social Networks* **34**(1), 101–111. cited By 14.

Lusher, D., Koskinen, J. and Robins, G.: 2012, *Exponential random graph models for social networks: Theory, methods, and applications*, Cambridge University Press.

Marco, A. C. and Rausser, G. C.: 2008, The role of patent rights in mergers: Consolidation in plant biotechnology, *American Journal of Agricultural Economics* **90**(1), 133–151.

Masrurul, M. M. et al.: 2012, An overview of strategic alliance: Competitive advantages in alliance constellations, *Advances In Management* .

McEvily, B. and Marcus, A.: 2005, Embedded ties and the acquisition of competitive capabilities, *Strategic Management Journal* **26**(11), 1033–1055.

McKelvey, M., Alm, H. and Riccaboni, M.: 2003, Does co-location matter for formal knowledge collaboration in the swedish biotechnology–pharmaceutical sector?, *Research Policy* **32**(3), 483–501.

Miller, D. J.: 2006, Technological diversity, related diversification, and firm performance, *Strategic Management Journal* **27**(7), 601–619.

Mowery, D. C., Oxley, J. E. and Silverman, B. S.: 1998, Technological overlap and interfirm cooperation: implications for the resource-based view of the firm, *Research policy* **27**(5), 507–523.

Narula, R. and Hagedoorn, J.: 1999, Innovating through strategic alliances: moving towards international partnerships and contractual agreements, *Technovation* **19**(5), 283–294.

Nelson, R. R.: 1990, On the complex economics of patent scope, *Colombia Law review* .

Nelson, R. R. and Winter, S. G.: 1982, The schumpeterian tradeoff revisited, *The American Economic Review* **72**(1), 114–132.

Nesta, L.J.J & Saviotti, P.: 2005, Coherence of the knowledge base and the firm's innovative performance, evidence from the us pharmaceutical industry, *Journal of industrial economics* **53**, 105–124.

Newman, M. E. and Girvan, M.: 2004, Finding and evaluating community structure in networks, *Physical review E* **69**(2), 026113.

Nieto, M. J. and Santamaría, L.: 2007, The importance of diverse collaborative networks for the novelty of product innovation, *Technovation* **27**(6), 367–377.

Niosi, J. and Zhegu, M.: 2005, Aerospace clusters: local or global knowledge spillovers?, *Industry & Innovation* **12**(1), 5–29.

Nonaka, I.: 1991, The knowledge-creating company, *Harvard business review* **Best of HBR**.

Østergaard, C. R.: 2009, Knowledge flows through social networks in a cluster, comparing university and industry links, *Structural Change and Economic Dynamics* **20**, 196–210.

Özdemir, K. and Saygili, M.: 2009, Monetary pressures and inflation dynamics in turkey: Evidence from p-star model, *Emerging Markets Finance and Trade* **45**(6), 69–86. cited By 5.

Pattison, P. and Robins, G.: 2002, Neighborhood-based models for social networks, *Sociological Methodology* **32**(1), 301–337.

Pattison, P. and Wasserman, S.: 1999, Logit models and logistic regressions for social networks: Ii. multivariate relations, *British Journal of Mathematical and Statistical Psychology* **52**(2), 169–194.

Pavitt, K.: 1984, Sectoral patterns of technical change: Towards a taxonomy and a theory, *Science Policy Research* (13), 343–373.

Penrose, E. T.: 1959, The theory of the growth of the firm, 1959, *Cambridge, MA* .

Pippel, G.: 2013, The impact of r&d collaboration networks on the performance of firms: a meta-analysis of the evidence, *International Journal of Networking and Virtual Organisations* **12**(4), 352–373.

Podolny, J. M.: 2001, Networks as the pipes and prisms of the market1, *American journal of sociology* **107**(1), 33–60.

Polanyi, M.: 1966, The tacit dimension, *library of congress* .

Powell, W. W., Koput, K. W. and Smith-Doerr, L.: 1996, Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology, *Administrative science quarterly* pp. 116–145.

Prencipe, A.: 1997, Technological competencies and product's evolutionary dynamics a case study from the aero-engine industry, *Research policy* **25**(8), 1261–1276.

Price, S. and Nasim, A.: 1998, Modelling inflation and the demand for money in pakistan; cointegration and the causal structure, *Economic Modelling* **16**(1), 87–103. cited By 4.

Pyka, A.: 2002, Innovation networks in economics: from the incentive-based to the knowledge-based approaches, *European Journal of Innovation Management* **5**(3), 152–163.

Pyka, A. and Scharnhorst, A.: 2009, Innovation networks, *Innovation Networks: New Approaches in Modelling and Analyzing, Understanding Complex Systems, ISBN 978-3-540-92266-7. Springer-Verlag Berlin Heidelberg, 2009* **1**.

Quintana-García, C. and Benavides-Velasco, C. A.: 2008, Innovative competence, exploration and exploitation: The influence of technological diversification, *Research Policy* **37**(3), 492–507.

Requardt, M.: 2003, Scale free small world networks and the structure of quantum space-time, *arXiv preprint gr-qc/0308089* .

Robbins, H. and Monro, S.: 1951, A stochastic approximation method, *The annals of mathematical statistics* pp. 400–407.

Robins, G., Pattison, P. and Elliott, P.: 2001, Network models for social influence processes, *Psychometrika* **66**(2), 161–189.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D.: 2007, An introduction to exponential random graph (p*) models for social networks, *Social networks* **29**(2), 173–191.

Rogers, E. M.: 1982, Information exchange and technological innovation, *in: Devendra Sahal (ed.) "the transfer and utilization of technical knowledge"* pp. 105–123.

Rowley, T., Behrens, D. and Krackhardt, D.: 2000, Redundant governance structures: An analysis of structural and relational embeddedness in the steel and semiconductor industries, *Strategic management journal* **21**(3), 369–386.

Salavisa, I., Sousa, C. and Fontes, M.: 2012, Topologies of innovation networks in knowledge-intensive sectors: Sectoral differences in the access to knowledge and complementary assets through formal and informal ties, *Technovation* **32**(6), 380–399.

Salle, I. and Yıldızoğlu, M.: 2014, Efficient sampling and meta-modeling for computational economic models, *Computational Economics* **44**(4), 507–536.

Sampson, R. C.: 2007, R&d alliances and firm performance: The impact of technological diversity and alliance organization on innovation, *Academy of Management Journal* **50**(2), 364–386.

Savin, I. and Egbetokun, A.: 2016, Emergence of innovation networks from r&d cooperation with endogenous absorptive capacity, *Journal of Economic Dynamics and Control* **64**, 82–103.

Saviotti, P. P.: 2007, On the dynamics of generation and utilisation of knowledge: The local character of knowledge, *Structural change and economic dynamics* **18**, 387–408.

Schrader, S.: 1991, Informal technology transfer between firms: Cooperation through information trading, *Research policy* **20**(2), 153–170.

Schumpeter, J. A.: 1942, *Capitalism, socialism and democracy*, Routledge.

Shan, W., Walker, G. and Kogut, B.: 1994, Interfirm cooperation and startup innovation in the biotechnology industry, *Strategic management journal* **15**(5), 387–394.

Snijders, T. A.: 2001, The statistical evaluation of social network dynamics, *Sociological methodology* **31**(1), 361–395.

Snijders, T. A.: 2002, Markov chain monte carlo estimation of exponential random graph models, *Journal of Social Structure* **3**(2), 1–40.

Snijders, T. A., Pattison, P. E., Robins, G. L. and Handcock, M. S.: 2006, New specifications for exponential random graph models, *Sociological methodology* **36**(1), 99–153.

Stolwijk, C., Ortt, J. and den Hartigh, E.: 2013, The joint evolution of alliance networks and technology: A survey of the empirical literature, *Technological Forecasting and Social Change* **80**(7), 1287–1305.

Strauss, D. and Ikeda, M.: 1990, Pseudolikelihood estimation for social networks, *Journal of the American Statistical Association* **85**(409), 204–212.

Suzuki, J. and Kodama, F.: 2004, Technological diversity of persistent innovators in japan: Two case studies of large japanese firms, *Research Policy* **33**(3), 531–549.

Szulanski, G.: 1996, Exploring internal stickiness: imepdiments to the transfer of best practise within the firm, *Strategic Management Journal* **17**(Special issue), 27–43.

Tang, M.-J., Puah, C.-H., Affendy Arip, M. and Dayang-Affizzah, A.: 2015, Forecasting performance of the p-star model: The case of indonesia, *Journal of International Business and Economics* **15**(2), 7–12.

Teece, D. J., Rumelt, R., Dosi, G. and Winter, S.: 1994, Understanding corporate coherence: Theory and evidence, *Journal of Economic Behavior & Organization* **23**(1), 1–30.

Ter Wal, A. L.: 2013, The dynamics of the inventor network in german biotechnology: geographic proximity versus triadic closure, *Journal of Economic Geography* p. lbs063.

Tödtling, F., Lehner, P. and Kaufmann, A.: 2009, Do different types of innovation rely on specific kinds of knowledge interactions?, *Technovation* **29**, 59–71.

Tomasello, M. V., Napoletano, M., Garas, A. and Schweitzer, F.: 2013, The rise and fall of r&d networks, *arXiv preprint arXiv:1304.3623* .

Trappey, C. V., Wang, T.-M., Hoang, S. and Trappey, A. J.: 2013, Constructing a dental implant ontology for domain specific clustering and life span analysis, *Advanced Engineering Informatics* **27**(3), 346–357.

Tsai, W.: 2001, Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance, *Academy of management journal* **44**(5), 996–1004.

Van Der Pol, J., Rameshkoumar, J.-P., Virapin, D. and Zozime, B.: 2014, A preminary analysis of knowledge flows: The case of structural composite materials in aeronautics.

van der Valk, T., Chappin, M. and Gijsbers, G. W.: 2011, Evaluating innovation networks in emerging technologies, *Technological Forecasting & Social Change* **78**, 25–39.

Van der Valk, T., Moors, E. H. and Meeus, M. T.: 2009, Conceptualizing patterns in the dynamics of emerging technologies: The case of biotechnology developments in the netherlands, *Technovation* **29**(4), 247–264.

Venables, W. N. and Ripley, B. D.: 2013, *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

Verspagen, B. and Duysters, G.: 2004, The small worlds of strategic technology alliances, *Technovation* **24**(7), 563–571.

Veugelers, R.: 1998, Collaboration in r&d: an assessment of theoretical and empirical findings, *De economist* **146**(3), 419–443.

Virapin, D. and Flamand, M.: 2013, L'innovation dans les matériaux composites, quelle diffusion autour de quels acteurs?, P2i conference 2013.

Von Hippel, E.: 1987, Cooperation between rivals: Informal know-how trading, *Research policy* **16**(6), 291–302.

Wasserman, S.: 1994, *Social network analysis: Methods and applications*, Vol. 8, Cambridge university press.

Wasserman, S. and Pattison, P.: 1996, Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp, *Psychometrika* **61**(3), 401–425.

Watson, J.: 2007, Modeling the relationship between networking and firm performance, *Journal of Business Venturing* **22**(6), 852–874.

Watts: 1999, Networks, dynamics and the small world phenomenon, *American journal of sociology* **105**(2), 493–527.

Watts, D. J. and Strogatz, S. H.: 1998, Collective dynamics of 'small-world' networks, *Nature* **393**(4), 440–442.

White, H. C.: 1992, *Identity and control: A structural theory of social action*, Princeton University Press.

Wuyts, S., Colombo, M., Dutta, S. and Nooteboom, B.: 2005, Empirical tests of optimal cognitive distance, *Journal of economic behavior & organization* **58**(2), 277–223.

Zaheer, A., McEvily, B. and Perrone, V.: 1998, Does trust matter? exploring the effects of interorganizational and interpersonal trust on performance, *Organization science* **9**(2), 141–159.